An Integrative Analysis of Alternative Polyadenylation

and MicroRNA Regulation in *Caenorhabditis elegans*

by

Stephen M. Blazie

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2016 by the
Graduate Supervisory Committee:

Marco Mangone, Chair
Joshua LaBaer
Douglas Lake
Stuart Newfeld

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

One of the fundamental questions in molecular biology is how genes and the control of their expression give rise to so many diverse phenotypes in nature. The mRNA molecule plays a key role in this process as it directs the spatial and temporal expression of genetic information contained in the DNA molecule to precisely instruct biological processes in living organisms. The region located between the STOP codon and the poly(A)-tail of the mature mRNA, known as the 3′Untranslated Region (3′UTR), is a key modulator of these activities. It contains numerous sequence elements that are targeted by *trans*-acting factors that dose gene expression, including the repressive small non-coding RNAs, called microRNAs.

Recent transcriptome data from yeast, worm, plants, and humans has shown that alternative polyadenylation (APA), a mechanism that enables expression of multiple 3′UTR isoforms for the same gene, is widespread in eukaryotic organisms. It is still poorly understood why metazoans require multiple 3′UTRs for the same gene, but accumulating evidence suggests that APA is largely regulated at a tissue-specific level. APA may direct combinatorial variation between *cis*-elements and microRNAs, perhaps to regulate gene expression in a tissue-specific manner. Apart from a few single gene anecdotes, this idea has not been systematically explored.

This dissertation research employs a systems biology approach to study the somatic tissue dynamics of APA and its impact on microRNA targeting networks in the small nematode *C. elegans*. In the first aim, tools were developed and applied to isolate and sequence mRNA from worm intestine and muscle tissues, which revealed pervasive tissue-specific APA correlated with microRNA regulation. The second aim provides genetic evidence that two worm genes use APA to escape repression by microRNAs in the body muscle. Finally, in aim three, mRNA from five additional somatic worm tissues was sequenced and their 3′ends mapped, allowing for an integrative study of APA and microRNA targeting dynamics in worms. Together, this work provides evidence that APA is a pervasive mechanism operating in somatic tissues of *C. elegans* with the potential to significantly rearrange their microRNA regulatory networks and precisely dose their gene expression.

DEDICATION

To my family, especially Marty and Deane Blazie, who provided just the right ingredients for my success.

To my wife, Nicole. We have endured quite a journey together. I really enjoyed the adventure and I am thankful to have done this with you by my side. I will dearly miss the fun we had exploring Arizona and making new friends. I am appreciative of your tremendous emotional and physical support during my years as a student in Arizona. I'm especially grateful for all the late night trips to the lab and having dinner ready for me after every long day. I love you! I could not have done it without you.

ACKNOWLEDGMENTS

My research advisor and mentor, Dr. Marco Mangone. You’re a superior example of what a mentor should be. I’ve been blessed with an incredible experience learning from someone who shares, and often surpasses, my passion for science. Your contagious excitement for basic biology always inspires me and you constantly motivated me toward unexpected directions that supported my growth as a scientist. I’m fortunate to have had the opportunity to work on the fascinating mechanism of alternative polyadenylation with you and I hope my contributions will provide a solid foundation for your research in the years ahead.

Dr. Josh LaBaer, my research advisor. I was lucky to have been an active member of your department. Having been under your umbrella, I had opportunities to receive your advice on a regular basis. It was almost like having the benefit of two mentors! I really appreciate your constant support!

Dr. Doug Lake, my research advisor. I appreciate your support to my graduate career through my years at ASU. My experience in your laboratory, while brief, provided me with skills that carried forward into my dissertation research. Ultimately, these biochemical tricks translated into the development of methods that kept on giving throughout the years. Thank you!

Dr. Stuart Newfeld, my research advisor. Your contributions to my graduate research and my development as a scientist have helped me enormously. I am thankful for your support through the years!

Justin Wolter, I was fortunate to have you as my science ‘brother’ growing up in the same lab family. It was especially great having someone who shares similar scientific interests around to discuss the day-to-day nuances of my research. You were often a source of substantial intellect in those discussions! Also, I appreciate your role as the official Mangone Lab social chairman and music aficionado, which made the lab a fun place to be. Your upbeat personality and positive demeanor always abated the stressful times. I wish you the best of luck in your career!

Kasuen Kotagama, It was a pleasure having someone with such character in the Mangone Lab! I really appreciate all of the help you’ve given me, be it with an extra set of hands on experiments or your insight during discussions. Your impressive attention to detail has come in handy more than once. You have all the makings of a great scientist. I hope I have not left you with too big of a mess! Good luck with the rest of your graduate career.

To all current and past members of the Mangone Lab, you have all sharpened my expertise and made me into better scientist. Thank you to former undergraduates Cherie Lynch, Henry Wilky, and Karina Ramirez for your tireless support with experimental design, troubleshooting, and execution. I wish you all the best of luck in your careers. Thanks especially to Alex Pierre-Bez, Cody Babb, and Victoria Godlove for your unlimited support with experiments and discussions!

Part of the fun of research is always meeting new and interesting people along the way! Thank you to all members of the Biodesign Institute who have supported my graduate research, provided useful advice, and collaborations.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

Genetics research facilitated by the small nematode C. elegans

Model organisms have provided an indispensible tool for understanding gene function in living organisms since the early days of genetic research. Over a century ago, Thomas Hunt Morgan's study of mutant genes using the fruit fly *Drosophila melanogaster* paved the way for discovery at the interface of genetics and development in multicellular species [1]. A key feature of effective model organisms for developmental genetics research is their genetic similarity to humans and the applicability of discoveries to improvements in human health. Mouse and other mammalian models are close representations of human physiology making them broadly useful for disease research. However, their organs and tissues are exceedingly complex and they require long periods of time to develop from embryo to adult, among other limitations. These disadvantages have argued in favor of simpler model systems to investigate many of the fundamental biological research questions.

Invertebrate models, such as the small, free-living soil nematode *Caenorhabditis elegans*, overcome many of these disadvantages. Sydney Brenner first adapted *C. elegans* in 1974 to begin long-term studies on how genes build a nervous system, a particularly challenging biological question to address in higher multicellular organisms. Dr. Brenner lauded their rapid and precise development (~3.5 days), limited and invariant number of somatic cells, and ease of basic genetic manipulation [2]. Further, *C. elegans* are primarily self-propagating hermaphrodites with the genotype XX (approximately 99.8% in population), where XO males (<0.2%) result from relatively rare non-disjunction events in meiosis. This feature allows experimentalists to execute genetic crosses, yet makes them easy to propagate in a laboratory setting [3].

In proceeding years, science has come to further appreciate *C. elegans* for its transparent tissues, simple methodology for transgenesis and its feasibility for genetic screens. Its relatively compact genome was also of the first sequenced metazoans, and the gene models have since been extensively mapped and characterized. The protein-coding portion of its genome shares ~70% homology to that of human, underscoring its applicability to modeling human disease. Remarkable discoveries in neurobiology, behavior, cell biology and development that have been translated to human biology are a testament to this feature. Notably, three such discoveries made in *C. elegans* laboratories have been awarded with the Nobel Prize in Medicine or Chemistry.

C. elegans is also an excellent model system to study how genes control somatic tissue development. They are eutelic organisms where every adult hermaphrodite has exactly 959 somatic cells [4]. Nearly all of these cells are transparent making them easy to visualize using standard light microscopy approaches. These unique features have enabled researchers to closely observe and map the organism's entire cell lineage from embryo to adult [5], a seminal undertaking that uncovered the widely conserved cellular mechanism of apoptosis. Taking advantage of forward genetic screens, scientists have coordinated this lineage map with the genes required for its precise development [6, 7]. These studies precisely mapped genetic pathways that pattern major worm organs such as the hermaphrodite vulva [8], the pharynx [9], and the male tail [10]. Importantly, the same genetic pathways play a fundamental role in mammalian development and many are implicated in human disease [11-13].

The basic anatomy of *C. elegans* shares many similarities with humans, having organ systems and tissues that are similar across metazoan species. This similarity makes them ideal for use in research aimed to investigate how these tissue structures are formed, since they are much simpler than mammalian tissues. For example, the human brain is excessively complex with over 100 billion neuronal cells formed with a network of synaptic connections, making it extraordinarily challenging to map neural connections and processes[14]. The eutelic aspect of *C. elegans* development has made it possible to map the connectivity of the exactly 302 neurons formed in adult hermaphrodites [15, 16]. Further, its genetic manipulability has made it possible to identify genes responsible for these connections [17]. Importantly, many of the genes identified so far have close homologs in human that share the same function. Therefore, basic research of tissue and organ biology in worms will allow us to identify genes and their networks that direct these developmental processes in metazoans.

C. elegans is also an effective model system to study human disease and genetic disorders. This is owed to the fact that the basic genetic mechanisms controlling the developmental events that establish cell fate are well conserved in metazoans. Artificial manipulation of these pathways in worms commonly emulates the human disease phenotype associated with those changes. Duchenne Muscular dystrophy (DMD) is an example of such a disorder. In humans, DMD is caused by a loss-of-function mutation in the dystrophin gene that encodes a structural protein important for muscle maintenance. The loss-of-function dystrophin allele causes a progressive loss of muscle activity in youth, eventually leading to paralysis [18]. This progressive deterioration in muscle activity is recapitulated in worms by introducing the same mutation in the worm dystrophin ortholog *dys-1*, suggesting that the same mechanisms leading to loss of muscle function are also conserved [19]. Therefore, the expansive genetics toolset available in worms is useful to study the precise series of molecular events underlying disease pathologies.

Many of the cellular processes that are commonly misregulated in cancers are also conserved in *C. elegans* [20]. For example, a program of controlled cell death, called apoptosis, is essential in normal states for regulating precise organ patterning in development [21]. In mammals, apoptosis is a powerful mechanism of tumor suppression that is commonly deactivated in cancer. The genetic components that direct apoptosis are also conserved in worms [22], allowing the study of factors that may influence these pathways in disease. Similarly, many oncogenes that drive cancers, such as Ras and Lin-28, are conserved in sequence and function in *C. elegans* [23, 24].

In summary, *C. elegans* have many suitable features that justify its use for questions in basic genetics that can also be translated to study human disease. In particular, worms are suitable for large-scale genetic screens that are generally not feasible in mammalian models. Such approaches can identify mechanisms that are less likely to be restricted to a particular biological context allowing general rules to be discovered. Finally, the experiments are done within living organisms having tissues, organs, and developmental stages where the mechanisms more closely represent what is occurring in these contexts.

Regulation of gene expression in metazoan development

In metazoans, development relies on the precisely coordinated activities of genome-encoded proteins in the proper time and space to direct mechanisms that specify cell fates. The regulation of gene expression is widely recognized as a key factor involved in this process. Mechanisms controlling gene expression involve multiple, often competing factors that function at every step of the central dogma: epigenetic (regulation of DNA accessibility), pre-transcription (regulation of RNA synthesis), post-transcription (regulation of RNA activity or protein synthesis), and post-translation (regulation of protein stability or activity), (**Figure 1.1**). While there are numerous examples of regulation occurring at each step, the precise mechanisms involved are not fully understood and novel modes where gene expression is controlled continue to be discovered. Apart from a few cases, it is also not understood exactly how controlling the dosage of gene expression drives tissue development forward.

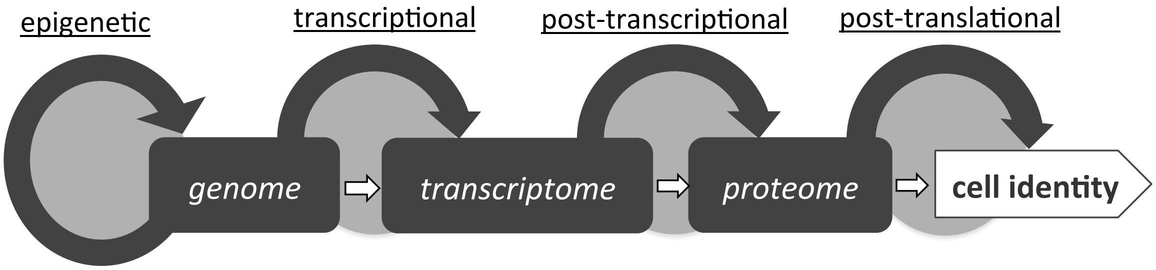


Figure 1.1. A complex program of gene regulation shapes cell identity. This illustration depicts the major modes of regulation (light gray circles) that precisely dose gene expression at each step of the central dogma (dark gray boxes), eventually culminating in tissue identity.

Perhaps the most famous example of gene regulation is that of X-chromosome inactivation in mammals. During female embryonic development, one allele of the X-chromosome is chosen for silencing, allowing somatic cells to achieve the correct dosage of gene expression from the remaining active X. This process also illustrates the often complex nature of the mechanisms that control gene expression in all metazoans. The overall process involves a combination of *cis* and *trans*-acting factors, primarily derived from genes encoded at the X-inactivation center. These include long non-coding RNAs that aide in the decision of which X-allele to inactivate and another that coats the chosen X-allele, recruiting with it numerous repressive protein complexes, including the polycomb repressor complex (PRC), that heavily methylate the chromatin of the chosen allele thereby repressing its expression [25]. X-inactivation is a notable example of a gene regulatory process requiring multiple mechanisms: those acting at the pre-transcriptional level (factors that direct expression of the lncRNAs) and others at the epigenetic level (ex. the PRC).

At the pre-transcriptional level, gene expression is often regulated by the activities of proteins called transcription factors that bind DNA sequences, usually in promoter or enhancer regions, and *trans*-activate the initiation of RNA synthesis [26]. Many transcription factors activate multiple genes in a network to strongly induce cellular events associated with the function of the target genes [26]. These proteins play a key role specifying cell fates during development [27] and their expression level often dictates the strength of their *trans*-activation on target genes. One of the most notable examples of such transcription factors are those belonging to the well conserved Hox gene family, which play an important role in tissue patterning in all metazoans [28]. Most of the Hox genes in *C. elegans* and other organisms are arranged on the chromosome in a single cluster where their expression pattern along the anterior-posterior body axis matches their position on the chromosome, a phenomenon termed co-linearity [29]. Their relative expression levels specify differential cell fates along the axis by activating the transcription of genes required for each unique identity [30]. Importantly, transcription factors like Hox are often expressed in gradients where precise expression levels substantially modulate their activities and the resulting cellular phenotypes.

It is therefore not surprising that many of the factors that control gene expression are often misregulated in disease. For example, misregulation of transcription factors, by deregulation of their expression or by mutation often misroutes cells to alternate, abnormal fates [31]. A notable example is the well-conserved transcription factor TP53, or p53. In normal states, this protein reacts to multiple cellular stressors including lack of nutrients or DNA damage [32]. Upon activation, p53 *trans*-activates multiple genes that prevent cell proliferation, activate DNA repair, or induce apoptosis, highlighting its role as a tumor suppressor protein [32]. In its absence, cells are subject to adverse effects from stressors and prone to adapt malignant cell fates. Accordingly, deregulated alleles of p53 are common in human cancers and some mutant alleles also have demonstrated oncogenic activities [33]. The example of p53 underscores the importance of factors that regulate gene expression in deciding cell fates.

Tissue identity is not conferred solely by a unique set of expressed genes. Rather, tissue-specific gene expression coupled with the precise dosing of these genes, coordinate activities that specify cell fate (**Figure 1.1**). Therefore altering not only absolute gene expression, but also their level of expression can have a dramatic impact on cell identity. Although many factors that dose gene expression have been characterized, it is becoming clear that other unexplored modes of gene regulation may exist [34]. Accumulating evidence suggests this may be especially true at the post-transcriptional level [35].

Post-transcriptional gene regulation induced by miRNAs

The messenger RNA (mRNA) molecule is a unique carrier of genetic information in that it is not restricted to the nucleus, allowing its interaction with a vast array of cellular factors. These interactions represent opportunities for gene regulation for example by sequestration of the mRNA [36], directing its localization [37], destabilization [38], degradation [39] or interfering with its translation [40]. A lot of this regulation is driven by RNA-binding proteins and non-coding RNAs that interact with small, mostly uncharacterized, cis-elements located in the three prime untranslated region (3′UTR) of the mature mRNA molecule (**Figure 1.2**). A particularly well-studied class of these *trans*-regulators, are the small ~22nt RNA molecules, called microRNAs (miRNAs) that repress gene expression post-transcriptionally through direct interactions with the mRNA.

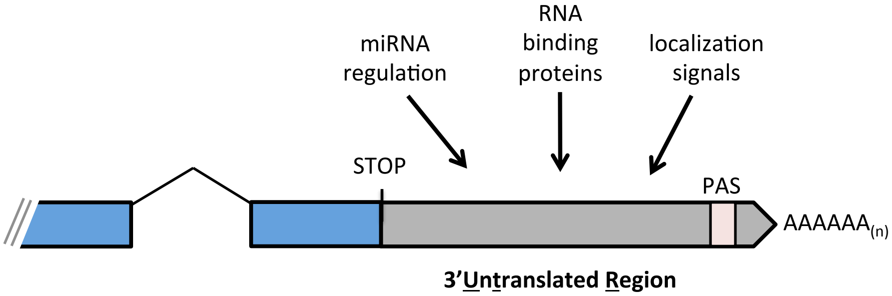


Figure 1.2. The 3′Untranslated Region. The 3′end of a representative eukaryotic gene model showing the terminal exons (blue) and the 3′UTR (gray). The 3′UTR contains numerous sequence elements that are typically targeted by miRNAs and RNA-binding proteins that dose gene expression or direct localization of the mature transcript.

The Victor Ambros laboratory is credited with the first reported discovery of a miRNA gene almost 30 years ago. They were intensively studying the heterochronic gene pathway, which controls precise transitions through larval development [41]. The Ambros lab was focused on a heterochronic gene, called *lin-4*, known to negatively regulate the levels of *lin-14* and coordinate transitions through the first larval stage. In their seminal paper, they report the lack of any functional open reading frame in the *lin-4* gene, indicating it did not encode a protein. They demonstrate that *lin-4* instead encodes a small, 22nt RNA with complementarity to sequences inside the 3′UTR of its target gene *lin-14*. Gary Ruvkun's laboratory then showed that sequences complementary to *lin-4* in the 3′UTR of *lin-14* are required for its post-transcriptional gene regulation [42].

The *lin-4* gene remained the only known miRNA up until seven years later when Gary Ruvkun's lab described that another heterochronic gene, *let-7*, encodes a miRNA that controls developmental timing by repressing expression of five genes at the post-transcriptional level [43]. Their work further demonstrated that the deletion of *let-7* induces a reversion of adult cells to larval fates and accordingly, its overexpression results in pre-mature induction of adult cell fates. The activities of *lin-4* and *let-7* are a reflection of what is appreciated about miRNAs today: they are capable of controlling dramatic developmental decisions in living organisms. The widespread conservation of *let-7* across multiple phylogenies including humans was reported shortly after its initial discovery [44], which prompted a concerted investigation into miRNA biology in numerous other species. Thousands of miRNA genes have since been identified in diverse species ranging from plants to viruses [45].

miRNAs are encoded from intergenic, intragenic, or intronic chromosomal loci and sometimes their genes are clustered [46]. All miRNA genes are transcribed into a long precursor RNA molecule, termed the primary miRNA (pri-miRNA) by RNA polymerase II. The polyadenylated pri-miRNA molecule forms a hairpin with a distinct stem-loop secondary structure that licenses it for processing by RNAse enzymes Drosha and Pasha [47]. The resulting pre-miRNA molecule is then exported from the nucleus by Exportin 5 where the hairpin is further cleaved by the RNAse Dicer into the mature ~22nt duplex structure in the cytoplasm. One of the two small RNA strands is then loaded onto Argonaute, the RNA-binding protein component of the RNA-induced silencing complex (RISC), by a still poorly understood mechanism.

Argonaute uses the loaded miRNA as a guide to target the RISC to small elements, usually located in the 3′UTRs, of target genes [45, 48]. The pairing of the miRNA loaded Argonaute can result in several outcomes, depending on the species and the degree of complementarity of the miRNA to the target mRNA [45]. The most common outcome of the pairing in metazoans is destabilization of the mRNA[49] [50], for example by recruiting deadenylases [51]. Inhibition of mRNA translation has also been reported [45]. Argonaute can also induce cleavage of the target mRNA in cases where complementarity to the miRNA is high [45].

A key challenge associated with the study of miRNA function in cells is the general lack of information regarding which genes they target [52]. Experimental approaches designed to identify targets are often laborious or lack the throughput required to identify biological targets from the ~20,000 annotated protein-coding genes in metazoans. Nonetheless, a few experimental approaches have been designed to identify precise miRNA targets in high-throughput, although they require a clone library of 3′UTR candidates to query [53],[54].

Computational approaches that search for likely biologically relevant miRNA targets have aided much of the miRNA research to-date. These approaches use weighted prediction algorithms that account factors known to be important for targeting, such as target complementarity and evolutionary conservation[45]. Central to these algorithms is the consideration that the nucleotides at positions 2-7 at the 5′ end of the miRNA, called the seed, are most often complementary to the target mRNA [45]. However, it is widely recognized that many miRNAs pair targets with imperfect complementarity, which may be supplemented by nucleotide pairing at the 3′end of the miRNA to facilitate targeting [55]. These factors are generally incorporated in the contemporary algorithms used for miRNA target prediction. Many such databases to assist with miRNA target prediction are now available [56] ,[57], [58], and their algorithms are constantly improved to reflect new discoveries in miRNA biology.

In terms of function, miRNAs are now appreciated for their key roles in development, in shaping gene expression and conferring robustness to biological pathways [59]. Their potential roles in forming and shaping gene expression gradients have put them at the forefront of research into post-transcriptional gene regulation. A pivotal bioinformatics study found that miRNAs are often expressed in tissue regions proximal to their targets, but are rarely co-expressed in the same cells [60]. This work suggests that miRNAs may play a role in shaping gene expression borders by clearing unneeded mRNA transcripts. Importantly, this work highlights the potential role of miRNAs in driving cell fate decisions within developing organisms. Further, disruption of overall miRNA function in metazoans gives rise to potent phenotypes. Deletion of the miRNA biogenesis factor Dicer altogether disrupts cell differentiation in mice [61] and results in severe germline defects and sterility in *C. elegans*.

miRNAs as a class are powerful regulators of gene expression. However, functions of most individual miRNAs in worms appear to be subtle. Although the first miRNAs described in worms are potent regulators of developmental timing [41],[43], the specific functions of most of the ~150 miRNAs identified in worms since then are unknown. An early study of 86 miRNA deletion mutant strains found that most of their miRNAs do not give rise to obvious disruptions in development or viability [62]. However, close examination of their activities in specific tissues revealed pervasive roles for several of them in development processes that are not crucial for overall viability. For example, *miR-61* is important for coordinating a feedback loop involving the Notch pathway in specifying vulva cell fates [63]. Similarly, *lsy-6* directs neuronal cell fate, specifying left-right asymmetry early in the development of a pair of head neurons [64]. These studies in worms highlight the importance of miRNAs in directing tissue-specific developmental programs.

Many aspects of miRNA biology are still not well understood and warrant further investigation. What are the targets and functions of individual miRNA genes? What factors control when and where a particular miRNA is able to actively target a gene? As miRNAs have been suggested to repress residual mRNA transcripts[59, 60], how may target genes similarly escape residual miRNAs from earlier developmental processes?

3′end formation of eukaryotic mRNA

The precise location of the 3′end of the mature mRNA molecule is important because it determines the landscape of its 3′ untranslated region (3′UTR) [65]. Since this sequence region contains many *cis*-elements that are targets for a variety of factors that dose gene expression, the mechanisms that direct the precise location and formation of the 3′end have lately become a subject of intense study. RNA transcription has been studied for more than half a century, focusing mainly on the initiation and elongation processes [66, 67]. However, we still do not understand the mechanisms of transcriptional termination, preparation of 3′ mRNA ends, or the regulation of these events that in turn control gene expression at a post-transcriptional level. Although the termination of transcription requires many of the same sequence elements as 3′end formation of the pre-mRNA, it is not yet clear exactly to what extent transcription termination influences 3′end choice and related processing [68].

Evidence from experiments in *Saccharomyces cerevisiae* has accumulated in favor of two mechanistic models for mRNA transcription termination in eukaryotes. In the 'torpedo model', the 5′ to 3′ exoribonuclease protein Rat1 is recruited to the RNA pol II CTD-tail along with 3′end processing factors. After cleavage of the transcript by 3′end formation ribo-endonucleases, Rat1 begins to degrade the extending transcript until it contacts and destabilizes RNA pol II, thereby terminating further transcription [69] [70]. In the second model, termed the 'allosteric' model, the elongation complex encounters sequence elements near the 3′end that induce a conformational change leading to destabilization and termination of transcription [68], [71]. Emerging evidence suggests these models are not mutually exclusive [72] and that both mechanisms rely on factors that induce cleavage and polyadenylation.

Cleavage and polyadenylation directs the processing of mRNA 3′ends, defining the precise end of the transcript and determining its 3′UTR landscape. This mechanism occurs co-transcriptionally, and for all mRNAs except for replication dependent histone encoding genes, is known to require at least two sequences near the 3′end of the nascent transcript (**Figure 1.3**). The poly(A) signal element (PAS) is a hexanucleotide located

~10-30 nucleotides upstream of the cleavage site that is most frequently the sequence 'AAUAAA' [73]. This sequence is necessary and sufficient for 3′end polyadenylation [74]. Notably, high-throughput sequencing efforts targeting the 3′end of transcripts have identified a larger array of PAS elements than initially indicated [75] suggesting they may modulate the efficiency of cleavage and polyadenylation and in turn regulate gene expression. Downstream of the cleavage site, another GU- or U-rich sequence is also involved [76] (**Figure 1.3**).

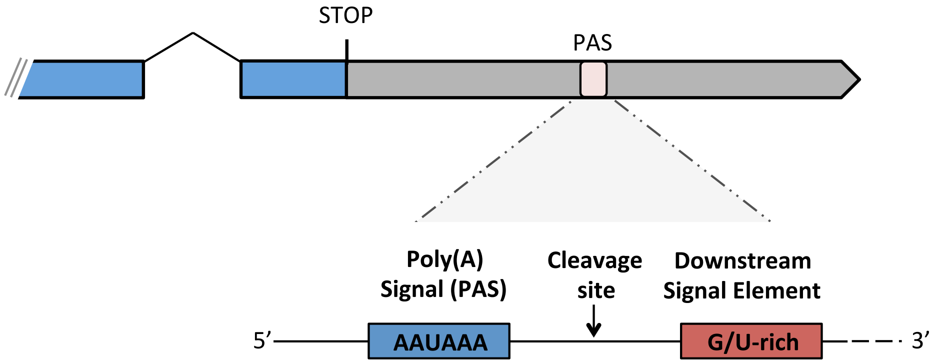


Figure 1.3. Sequence elements at the poly(A) site direct 3′end formation. In the eukaryotic 3′UTR (gray), the poly(A) site (pink box) contains a hexameric poly(A) signal element (blue box) and a downstream G/U-rich element (red box) that direct factors to the cleavage site (arrow) to induce 3′end cleavage and poly(A)-tail addition.

These sequences facilitate recruitment of two large multimeric complexes named Cleavage and Polyadenylation Specificity Factor (CPSF) and Cleavage Stimulation Factor (CstF), [77-79]. CPSF recognizes and binds to the PAS element located ~19nt from the poly(A) site in the 3′ UTR of mRNAs. CstF directly interacts with CPSF and binds to the GU-rich sequence downstream of the cleavage site. Together, these large complexes remodel the local RNA structure and recruit endonucleases (CFI and CFII) that cleave the mRNA and poly(A)-polymerases that produce the poly(A)-tail [80] (**Figure 1.4**).

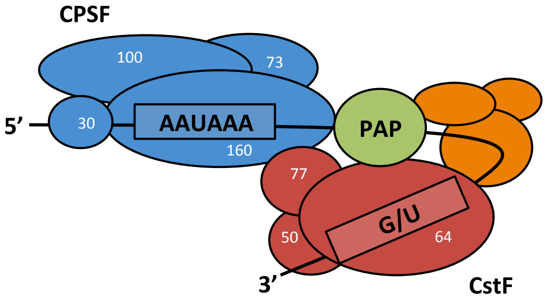


Figure 1.4. CPSF and CstF direct 3′end formation events. The multimeric CPSF (blue) and CstF (red) complexes bind to the hexameric PAS element 'AAUAAA' and G/U-rich downstream element, respectively. The complexes remodel the local mRNA structure and facilitate recruitment of endonucleases (orange) that cleave the transcript and poly(A)-polymerase (green) that adds the poly(A)-tail.

Several other factors involved in cleavage and polyadenylation have been identified in a series of biochemical experiments performed over the past ~25 years [80] leading to the idea that mechanisms coordinating formation of the mRNA 3′end may be much more complex than once appreciated. Many of these factors may be involved in assembly of the polyadenylation machinery or maintaining its structure [81]. Others may stimulate its activity or efficiency. For example, U1-snRNP has been shown to directly interact with CPSF and positively influence its activity [82].

A recent biochemical study revealed that the human pre-mRNA 3′end formation complex is over 1 megadalton and contains over 50 proteins that have no previously described roles in 3′end formation, suggesting much more complexity and flexibility in function than what is currently appreciated [83]. Many of these factors are also conserved in worms [84], suggesting the existence of widespread mechanisms that regulate cleavage and polyadenylation.

Alternative polyadenylation

Recent transcriptome data from yeast [85], plants [86], mammals [87], and nematodes [88, 89] has shown that alternative polyadenylation (APA), a poorly understood mechanism in which the same gene is expressed with multiple 3′UTR isoforms, is pervasive in metazoans (**Figure 1.5**). Although this process could dramatically impact gene expression by controlling the *cis*-regulatory landscape of the 3′UTR, the mechanisms that direct APA events are still poorly understood. There are currently no general, unified mechanisms described for how the cleavage and polyadenylation machinery discriminates between multiple PAS elements in the same 3′UTR. However, several factors have been identified that influence poly(A) site choice and direct APA for a few cases.

In one such example, an RNAi screen in human cells for protein factors that drive APA identified the nuclear poly(A)-binding protein (PABPN1) as a regulator of this process[90, 91]. PABPN1 blocks usage of the proximal PAS sites, restricting cleavage events to the PAS most distal from the STOP codon and extensive preferred expression of long 3′UTR isoforms throughout the transcriptome. Further studies found an association between APA-induced 3′UTR shortening along with low expression levels of PABPN1 and poor prognosis in small-cell lung cancer [92]. This highlights an important association between mechanisms that control APA and disease.

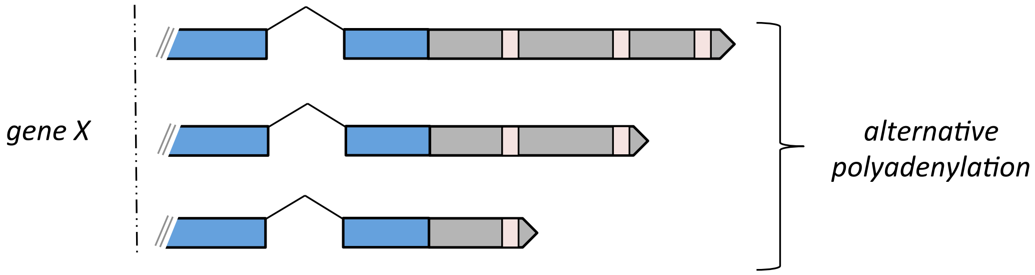


Figure 1.5. Alternative polyadenylation allows expression of more than one 3′UTR isoform per gene. In this hypothetical illustration, a gene (*gene X*) is expressed with multiple 3′UTR isoforms due to the presence of multiple poly(A) sites (pink boxes) in the same 3′UTR (gray).

A recent study has found that the U1 small nuclear RNA (U1 snRNA) similarly regulates APA, by recognizing and binding to upstream (proximal) poly(A) sites during transcription and blocking 3′end formation at these sites [93]. Depletion of U1-snRNP resulted in increased cleavage and polyadenylation at proximal PAS sites and expression of short 3′UTR isoforms. Interestingly, this work also showed a decrease of U1 snRNA levels in neuronal and immune cell activation. This result points to a potential link between APA mechanisms and the known shortening of 3′UTRs during induced cell proliferation (discussed below).

Recent systematic analysis found that extensive combinatorial variation between dosing of the core poly(A) machinery, splicing factors and the distance between poly(A) sites across developmental contexts are factors in determining poly(A) site choice [94]. These data enforce the idea that APA may be controlled by multiple combinatorial mechanisms.

Although it is widely anticipated that APA has a widespread impact on gene expression, the precise influence of the 3′UTR isoforms that result from APA are still not clear. 3′UTR isoforms containing distinct targets for factors that regulate gene expression may allow for the generation of a wide array of gene expression outputs in a context dependent manner. Several pivotal works have demonstrated the potential for APA to be regulated on a context-specific basis in support of this hypothesis.

Sandberg and colleagues showed that activation of primary mouse CD4+ T-lymphocytes results in global APA-induced shortening of their 3′UTRs [95]. This work also found a consistent correlation between proliferation and 3′UTR shortening among multiple tissue types, supporting the hypothesis that increased proliferation states are supported by a general release from 3′UTR mediated regulation through APA. Christine Mayr and colleagues extended these results in a seminal finding that global 3′UTR shortening is a characteristic of cancer cell lines [96]. Several oncogenes that exhibit this phenomenon in the tested cancer cells exclude miRNA targets and presumably avoid post-transcriptional gene regulation, suggesting a role for APA in supporting the oncogenic state.

While these studies establish a link between APA and miRNA target loss, they do not demonstrate a direct role for APA in allowing genes to escape regulation by the activities of co-expressed miRNAs. So far, only one known example of a specific APA event used to escape miRNA regulation to dose gene expression in support of function has been described. Pax3, a transcription factor controlling muscle cell differentiation, is tightly regulated in muscle stem cell tissue by the miR-206 miRNA. In a subset of these cells, Pax3 escapes regulation by miR-206, allowing its expression to levels required to drive muscle cell differentiation [97]. This example, coupled with the widespread and tissue-specific expression of APA, suggests that APA could dose gene expression by allowing their tissue-specific escape from miRNA regulation among many different cell types. Indeed, data from retina, brain, placenta and other human tissues [98-100] suggests APA functions extensively at a tissue-specific level.

Alternative polyadenylation in C. elegans

The general lack of comprehensive 3′UTR annotations for metazoan transcriptomes coupled with the complexity of most mammalian model tissues has made it challenging to identify general mechanistic rules for APA and to understand its activities in cells. Recently, two large-scale efforts have finely sequenced and annotated the 3′UTRs of almost every protein-coding gene in the *C. elegans* transcriptome [88, 89]. While both groups used very different approaches to both sequence and map worm 3′UTRs, their results are largely cross-validating and provide a gold-standard set of expressed 3′UTR isoforms for APA research. These results have provided insights into APA that point to opportunities to now study its mechanisms, its precise tissue-specific expression patterns, and its activities in regulating gene expression.

Both 3′UTRome mapping efforts confirm the prevalence of APA in worms. Almost half (~46%) of worm genes are expressed with multiple 3′UTR isoforms and approximately 8% of their genes have more than five 3′UTR isoforms. It is unclear why worm genes require so many different 3′ends, but its widespread nature in the worm transcriptome suggests it may broadly impact the regulation of gene expression through development and in specific tissues.

Indeed, distinct patterns of 3′UTR isoform expression through *C. elegans* development support this hypothesis. Mangone *et al.* prepared and sequenced developmental stage-specific mRNA libraries finding that average 3′UTR length decreases through worm age [88]. Embryos and dauer worms preferred longer 3′UTR isoforms, while adult worms predominate shorter 3′UTRs. These data suggest a tendency for increased regulation earlier in development and a release from 3′UTR mediated regulation in later stages coordinated by APA. Further, these data give support to the idea of context-specific regulation of APA to support temporal and perhaps also spatially unique gene expression programs.

The poly(A) signal elements used to induce 3′end formation events in worms have also been mapped and studied transcriptome-wide. This revealed a surprisingly infrequent use of the canonical ‘AAUAAA’ hexamer where only ~39% of 3′UTR isoforms use this PAS for cleavage [88, 89]. The remaining sites use PAS elements that differ from the canonical PAS by one or two nucleotides. Additionally, a surprising ~13% contain no detectable PAS element near the cleavage site suggesting unconventional mechanisms of 3′end formation may be used in these cases [88]. Interestingly, genes having two or more 3′UTR isoforms tend to use the canonical PAS to cleavage at sites that are distal from the STOP codon and variant hexamers or no PAS for cleavage at proximal sites. This result points to potential mechanisms that may drive preferred cleavage at proximal PAS sites to allow a release from 3′UTR mediated regulation. Although it is known that many human mRNA 3′ends are also sometimes cleaved with non-canonical PAS elements [75], the now comprehensive atlas of PAS element usage in the worm transcriptome makes it a unique resource to study this poorly understood process in metazoans.

Hypotheses and specific aims

APA likely provides a powerful regulatory mechanism using combinatorial variation between *cis*-elements and trans-acting factors to regulate gene expression in a tissue-specific manner (**Figure 1.6**).

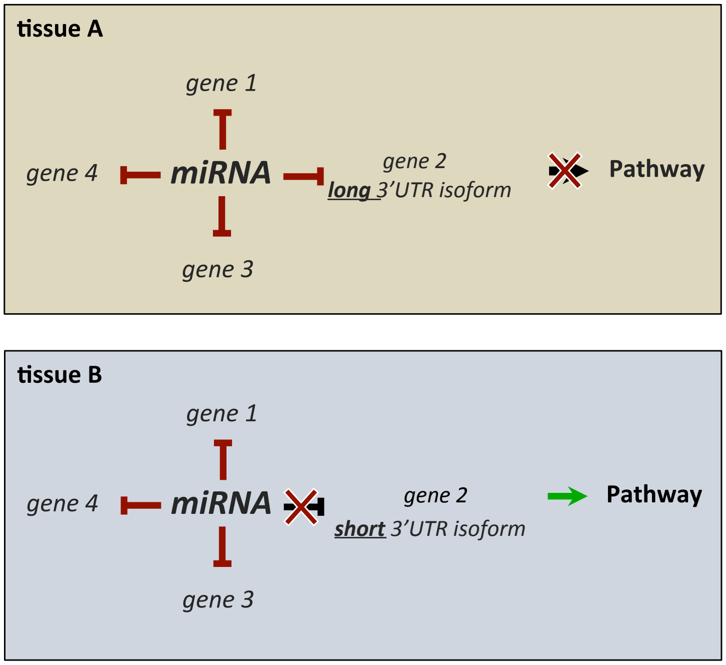


Figure 1.6. Working hypothesis: APA reorganizes miRNA regulatory networks at the tissue-specific level to direct their local activities. In this hypothetical example, tissues A (brown) and B (blue) express the same miRNA that targets genes 1 through 4. gene 2 is uniquely important for inducing a pathway required for tissue B identity and achieves the dosage required to carry out this function through an APA event specific to tissue B. The opposite APA event expresses the long 3′UTR isoform in tissue A, maintaining the miRNA target and blocking entry into the pathway.

Except for a few single gene anecdotes this idea has not been systematically explored. The broad questions addressed by this research are: 1) How pervasive is APA in living tissues? 2) What is its fundamental role? 3) How does 3′UTR heterogeneity integrate with negative regulatory networks driven by miRNAs to reshape gene expression in normal states? This research will 1) map APA dynamics at the tissue-specific level in living organisms, to learn patterns, rules and mechanisms using systems biology approaches, and 2) provide mechanistic insights and genetic validation on how APA interfaces with miRNA regulatory networks. Although several researchers use mammalian cell, or cancer cell systems to study APA, *C. elegans* has been selected for these experiments because when compared to other metazoans, its 3′UTRome is currently the best-annotated 3′UTR dataset available. *C. elegans* also possesses the best miRNA target predictions, and it has very few cells along with a finely annotated developmental lineage, enabling the study of APA and miRNAs crosstalk thorough development. In addition, worms are amenable for *in vivo* studies in an intact animal and can be easily manipulated with powerful genetic and molecular biology tools. Previous studies have used gene-by-gene approaches and unique contexts, such as cancer, that are both lacking in that the findings may not be applicable to the whole organism or tissue studied.

Moreover, APA is likely context dependent in terms of cell type and environment, which supports a genome-wide approach in *C. elegans*, where these contexts are already in place. The research presented here uses well-established techniques, such as RNA-IP, Next-Gen sequencing and others, that so far have rarely been applied to studying APA.

The first aim focuses on the development of biochemical methods to isolate and sequence tissue-specific mRNA from *C. elegans* intestine, pharynx, and muscle tissues. The sequence data will be used to bioinformatically map poly(A) clusters corresponding to the 3′ends of tissue-specific mRNA to identify intestine and muscle-specific 3′UTR isoform expression. These data will be used to closely study the extent and dynamics of intestine and muscle specific APA in worms and address the hypothesis that much of the 3′UTR heterogeneity observed in the worm 3′UTRome is restricted to specific tissues.

This approach will also allow mapping and identification of potential tissue-specific sequence elements located near the cleavage sites that coordinate tissue-specific APA events.

The second aim will develop and employ reporter-based genetics tools in transgenic worms to validate select genes having tissue-specific APA events that allow them to counteract co-expressed miRNAs. These experiments will provide the first evidence for worm genes that use APA to escape regulation by miRNAs to support their dosage to levels required to drive their local activities. This is in direct support of the overall research hypothesis that APA rearranges miRNA targeting networks in specific tissues.

Finally, the third aim will comprehensively investigate the hypothesis that ubiquitously expressed genes in *C. elegans* use APA to dose their expression levels uniquely in each tissue by modulating potential miRNA targeting events. This aim will apply the tissue-specific mRNA isolation strategies from aim one to five additional somatic worm tissues (hypodermis, seam cells, arcade cells, GABAergic neurons, and NMDA neurons) that cover most of the *C. elegans* somatic tissue anatomy. This approach is expected to 1) further refine the commonly expressed and tissue-restricted gene pools and 2) allow for an in-depth examination of the patterns of APA-induced miRNA target loss between tissues.

CHAPTER 2

COMPARATIVE RNA-SEQ ANALYSIS REVEALS PERVASIVE ALTERNATIVE POLYADENYLATION IN CAENORHABDITIS ELEGANS INTESTINE AND MUSCLES.

Publication Note

The research reported in this chapter was previously published in *BMC Biology*. Stephen M Blazie, Cody Babb, Henry Wilky, Alan Rawls, Jin G Park, and Marco Mangone. Comparative RNA-Seq analysis reveals pervasive alternative polyadenylation in *Caenorhabditis elegans* intestine and muscles. BMC Biology. 2015;13(1).

Overview

Transcriptome plasticity is a powerful modulator of gene expression that provides multicellular organisms with the complexity needed to drive development and maintain tissue identity. Apart from a few cases, we still do not fully understand what specific transcriptome rearrangements occur in somatic tissues, and how they integrate with gene regulation at the post-transcriptional level to maintain tissue and cellular diversity.

The small nematode *C. elegans* is an ideal model organism to study these events, since its gene model has been extensively characterized in past years [101]. It is also experimentally tractable with short and precise developmental timing, ~1,000 somatic cells, a transparent and simple body plan, and an entirely defined cell lineage [102]. Large-scale efforts have detailed its transcriptome at a global level [101]. Promoter diversity [103], alternate splicing events [104], and changes in 3′ untranslated regions (3′UTRs) [88, 89] are also well characterized.

While the formation of several worm tissues and the genes involved in driving these processes have been extensively described [105] [106], we still do not fully understand how the synergistic activity of tissue-specific events before, during, and after transcription drive and maintain tissue identity. Pre-transcriptionally, enrichment of sequence-specific elements within *C. elegans* promoters has been linked to tissue-specific changes in gene expression [107-109], suggesting that these elements, together with *trans*-acting factors that recognize them, are fundamental for driving the transcriptional programs of unique tissues.

During transcription, mRNA isoforms resulting from alternative splicing increase transcriptome complexity, coordinating tissue development [110]. Recently, a genome-wide study in *C. elegans* found that thousands of transcripts are alternatively spliced and many of them change splicing patterns during development [104], suggesting that tissue-specific splicing may play key roles in this process.

Post-transcriptionally, 3′UTRs are known to contain multiple regulatory sequence elements important for gene regulation [45]. Recently, two independent studies suggest that more than 40% of worm genes possess 3′UTRs subjected to alternative polyadenylation (APA), a mechanism that generates multiple 3′UTR isoforms for the same genes [88, 89]. This process is widespread in metazoans [96, 111], coordinated through development [88, 89], and misregulated in disease [96], underscoring a potential role for APA in tissue-specific modulation of gene expression.

The cleavage and polyadenylation of nascent mRNAs in eukaryotes is mainly executed by two large multimeric complexes named Cleavage and Polyadenylation Specificity Factor (CPSF) and Cleavage Stimulation Factor (CstF) [112]. CPSF recognizes and binds to the Polyadenylation [poly(A)] signal (PAS) element located ~19nt from the polyA site in the 3′UTR of mRNAs. In metazoans, the PAS sequence is commonly AAUAAA [112]. This sequence is necessary and sufficient for 3′end polyadenylation [112]. CstF directly interacts with CPSF and binds to GU-rich elements downstream of the cleavage site [112].

Although APA is pervasive in worms and correlated with development, suggesting that APA functions in worms tissues [88], it is unclear whether APA is tissue-specific.

Both CPSF and CstF are likely to have a role in managing the choice between PAS elements in the same 3′UTR and inducing APA. There may also be additional tissue-specific accessory factors that modify the basal polyadenylation machinery, controlling the usage of one PAS element over another. Tissue-specific isoforms of the CPSF or CstF complexes could be responsible for APA [113, 114]. Over a decade ago, stoichiometric levels of CstF members were indeed shown to control APA in B cell activation [115], and recent high-throughput approaches showed that other factors might also play important roles in modulating APA [116, 117]. Other processing factors were also recently shown to influence the location of cleavage [117]. These studies underscore the importance of the correct stoichiometric ratio of each of the 3′end processing factors for producing a mature mRNA. Surprisingly, it was also recently shown that U1 snRNP is involved in this process, suggesting possible cross talk between APA and the RNA splicing machinery [93]. These models may not be mutually exclusive.

In *C. elegans,* the isolation of tissue-specific mRNA to study transcriptome plasticity and APA is challenging due to the lack of *in vitro* cell cultures, the worm’s tough outer cuticle that interferes with sample preparation, and the small size of many tissues that prevents manual dissection. Several techniques have been developed to circumvent these issues, including fluorescence-activated cell and nuclear sorting [118, 119], nuclei-tagging [120], and mRNA-tagging [121]. In particular, mRNA-tagging has been widely used to isolate and study mRNA from muscle [122, 123], epithelial [124], hypodermal [125], neuronal [126] and seam cells [123]. This technique uses tissue-specific promoters to drive expression of a FLAG epitope-tagged cytoplasmic poly-A binding protein (PABPC), which specifically binds to the poly(A)-tail of mRNAs in the cytoplasm, followed by crosslinking and immunoprecipitation of tissue-specific mRNAs. Thus far, mRNA-tagging has been coupled with DNA microarrays and genomic tiling arrays, which both lack the advanced sensitivity and specificity possible today with deep sequencing, potentially limiting the results obtained with this methodology.

Recently, a novel method that couples tissue-specific nuclei isolation with deep sequencing was used to analyze the *C. elegans* intestine transcriptome [119]. Unfortunately, while enhancing the sensitivity of tissue-specific mRNA extraction and sequencing, this technique limits the analysis to nuclear mRNA species, and only identifies short 3′end portions of 3′UTRs that need to be bioinformatically attached to their closest genes, potentially introducing mapping inaccuracies. In addition, this method does not provide tissue-specific mRNA isoform data, losing an important component needed to study the transcriptome plasticity of individual tissues.

Integrating mRNA-tagging with RNA-Seq analysis could significantly improve the resolution of these studies and identify additional factors controlling tissue development and identity. Here, we have improved tissue-specific transcriptome profiling in *C. elegans*, optimizing mRNA-tagging for deep sequencing. We call this updated method polyA-tagging and sequencing (PAT-Seq). We have applied PAT-Seq and profiled tissue-specific mRNA from the *C. elegans* intestine and two muscle tissues belonging to the pharynx and body wall of mixed stage worms. We describe and compare gene expression, promoter sequence composition, mRNA isoforms changes, and alternative polyadenylation between each tissue.

PAT-Seq significantly improved the resolution of tissue-specific transcriptomes from previous studies, adding thousands of novel genes and isoforms that allow for a more comprehensive analysis of them. In addition, we have used PAT-Seq to profile the transcriptome of a previously uncharacterized tissue (pharynx), allowing us for the first time to directly compare gene expression changes between different tissues from the same organism at a higher resolution, using the same experimental settings.

We find that transcript diversity detected among these three tissues is mirrored by the presence of characteristic tissue-specific promoter signatures. In addition, we found that APA is widely used at a tissue-specific level, highlighting major complex tissue-specific transcript dynamics and post-transcriptional regulatory mechanisms. We describe a large number of 3′UTR isoforms specifically expressed in each tissue and find that these 3′UTRs are enriched for experimentally and bioinformatically predicted microRNA (miRNA) targets, suggesting that tissue-specific APA is used in worms as a mechanism to interface with miRNA mediated post-transcriptional gene regulation.

Finally, we have remapped, incorporated, and curated 3′UTR data from previously published studies [88, 89, 101, 119, 127] and integrated these data with our new tissue-specific datasets. The database is accessible through our new worm APA-specific website

[128], which is publically available and represents a unique resource for the scientists interested in 3′UTR biology.

Results

Isolation of mRNAs from intestine and muscle tissues

mRNA-tagging has so far been coupled with low-resolution platforms, such as microarrays and tiling arrays, that lack the sensitivity required to detect low expressed transcripts, pinpoint gene isoform changes, and map 3′UTRs at single base resolution. To improve upon its sensitivity and specificity, we made several key changes to the original mRNA-tagging protocol [121, 129]; 1) We added three tandem FLAG-epitope (3xFLAG) tags instead of one, to improve the efficiency of the FLAG pull-down [130]. 2) In the original mRNA-tagging protocol, the FLAG-tagged PABPC construct is selectively expressed as extra chromosomal arrays, which are unstable and often lead to mosaic expression patterns, or integrated as multiple copy lines [129, 130]. We instead opted for the widely used Mos-1 single copy insertion (MosSCI) technology [131], which stably incorporates the construct of interest in the worm genome. 3) We adopted a novel strategy to prepare the tissue-specific cDNA libraries that relies on linear amplification of mRNAs, minimizing the quantification error rate due to limited starting material, providing high-quality transcriptome and 3′end data in the same experiment [132], and 4) replaced the microarray step with Next Generation sequencing (Illumina HiSeq) to improve data resolution.

We used tissue-specific promoters to drive the expression of the *C. elegans* cytoplasmic PABPC gene (*pab-1*), in-frame with GFP and fused to a 3xFLAG tag (PolyA-Pull), in intestine, pharynx, or body muscle (**Figure 2.1**, see Experimental).

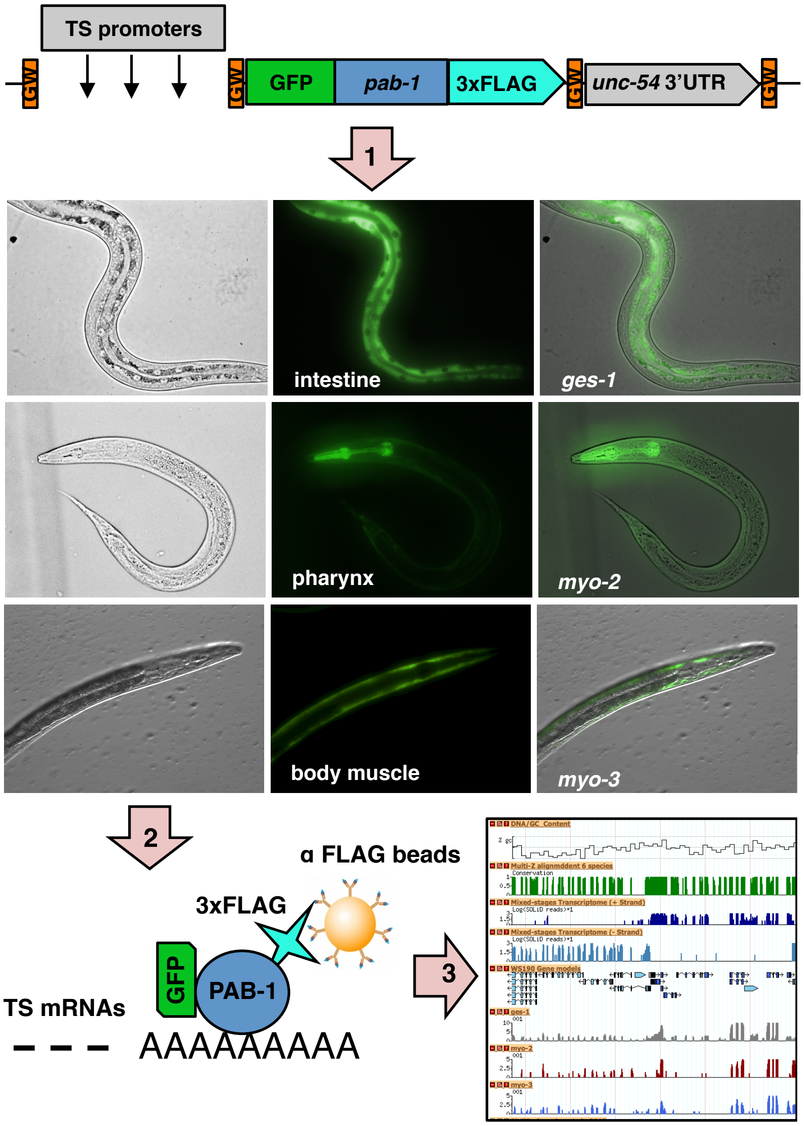


Figure 2.1. Overview of the PAT-Seq approach. PAT-Seq uses Gateway-compatible (GW) entry vectors expressing the PolyA-Pull cassette in each tissue using tissue-specific (TS) promoters. (1) PolyA-Pull expressed in the intestine (*ges-1* promoter), pharynx (*myo-2*), and body muscle (*myo-3*). (2) Expression of PolyA-Pull produces a 3 × FLAG-tag (light blue) fused to PAB-1 (blue), which specifically binds to the poly(A) tails of mRNAs (TS mRNAs). The complex is immunoprecipitated using α-FLAG beads. (3) Tissue-specific cDNA libraries are sequenced and mapped onto the WS190 gene model.

As MosSCI technology mediates transgene and rescue cassette (*unc-119*) insertion into a specified region of chromosome II, we confirmed this integration event in each transgenic worm line (**Figure 2.2, left panel**), and verified its expression using western blotting (**Figure 2.2, right panel**).

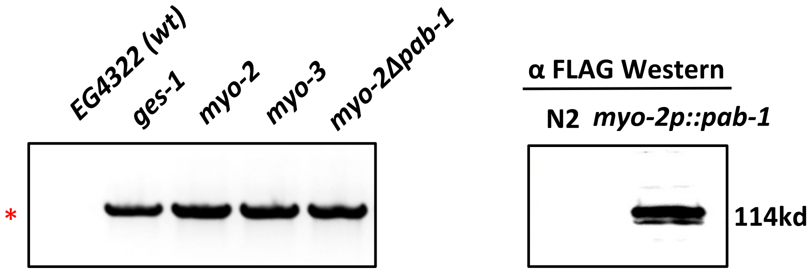


Figure 2.2. Detection of stable integration of the PolyA-Pull cassette. *Left panel*: Using PCR we detected genomic integration of the common portion of the PolyA-Pull cassette (2.6 kb band, red asterisk) in each tissue. The negative control, *myo-2Δpab-1*, was also integrated. *Right panel:* Western blotting using α-FLAG antibodies detected the in-frame PolyA-Pull fusion protein in lysates from transgenic worms expressing it in the pharynx (*myo2::pab-1*) but not in lysate from wild type N2 worms.

We tested the sensitivity and tissue-specificity of our mRNA pull-down approach using worms expressing PolyA-Pull in the pharynx (*myo-2p::PolyA-Pull*) (**Figure 2.2**), validating the ability of our construct to selectively bind muscle specific transcripts (**Figure 2.3, lanes 5 and 6**). The known intestine-specific transcript *ges-1* (**Figure 2.2**, lane 7) and hypodermis-specific *dpy-7* (**Figure 2.3, Lane 8**) were depleted from the same sample. Immunoprecipitation from wild type N2 worms yielded no detectable background in RT-PCR for the same genes (**Figure 2.3, lanes 9-12**). To test if our PolyA-Pull construct selectively binds polyA+ RNAs, we prepared a GFP::3xFLAG fusion protein that does not contain *pab-1*. This vector is unable to bind polyadenylated mRNAs (Δ*pab-1*-Pull, see Experimental). We expressed this new construct in the pharynx (*myo-2p*::Δ*pab-1-Pull*). As expected, using this construct, we were unable to detect the pharynx-specific *myo-2* transcript (**Figure 2.3, *right***).

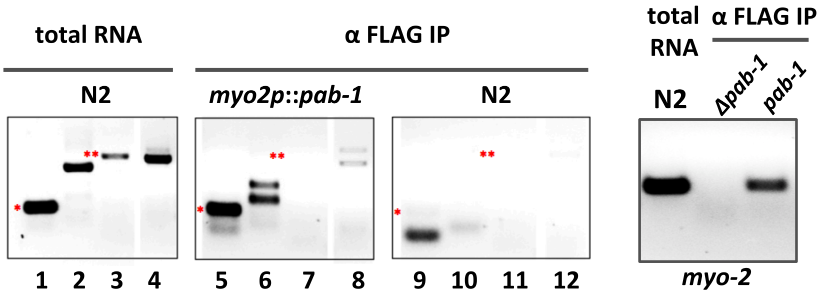


Figure 2.3. Quantification of the specificity and sensitivity of the pull-down using RT-PCR. *Left panel: myo-2* (lane 1) (*), *ges-1* (lane 3) (**) and *dpy-7* (lane 4) transcripts were detected in total RNA extracted from wild type N2 worms. Middle panel: Using immunoprecipitation, we successfully detected the presence of the muscle-specific gene *myo-2* (lane 5) (*) and the exogenous *unc-54* 3′UTR (lane 6), but not the intestine-specific *ges-1* (lane 7) (**) and the hypodermis-specific *dpy-7* (lane 8). These transgenic worms expressed PolyA-Pull cassette in the pharynx, but not in our negative control in wild type N2 worms (lanes 9-12). The primers used to detect *unc-54* 3′UTR also detected 18S rRNAs (lane 2). This band was replaced with two *unc-54* 3′UTR isoforms (lane 6), suggesting that PolyA-Pull enriched for polyA+ RNAs. Right panel: We are unable to isolate tissue-specific RNA from worms lacking *pab-1* (*Δpab-1*).

Taken together, these results suggest that our PolyA-Pull construct effectively enriches tissue-specific mRNAs and specifically binds polyA+ RNAs.

PAT-Seq analysis of mRNAs from intestine and muscle tissues

We then prepared our tissue-specific libraries with two biological replicates (6 preps in total). We also performed two negative control pull-down experiments expressing the Δ*pab-1*-Pull construct in the pharynx to optimize the sensitivity and specificity of our approach. Following the pull-down, the cDNA libraries were prepared using isothermal linear cDNA amplifications, which allows cDNA synthesis across the full length of the transcripts with as little as 1 ng of total RNA, thus improving the coverage of our whole-transcriptome amplification independently of the 3′polyA tail [132]. We then barcoded, pooled, and sequenced our eight RNA pull-down libraries on the Illumina HiSeq-2000 platform (see Experimental). The resulting paired reads were computationally assembled and mapped onto the *C. elegans* WS190 genome. A summary of the results of this mapping is displayed in **Table 2.1**.

We obtained ~15M unique reads per sample (~130 million reads total). The number of genes mapped in each tissue was consistent between biological replicates, reflecting the robustness of our library preparation from tissue-specific RNA samples.

samples (tissue)		total reads	mapped (%)	not mapped	average depth
intestine	*experiment*	18,148,228	11,473,900 (63%)	6,674,328	65.2x
	replicate	14,948,680	9,376,501 (63%)	5,572,179	49.6x
pharynx	*experiment*	15,685,384	12,028,022 (77%)	3,657,362	17.4x
	replicate	14,798,098	10,370,652 (70%)	4,427,446	28.5x
body muscle	*experiment*	15,496,850	10,818,093 (70%)	4,678,757	9.8x
	replicate	16,885,324	12,829,775 (80%)	4,055,549	12.1x
myo-2Δpab-1 (- control)	*experiment*	13,644,473	10,589,072 (78%)	3,055,401	51.2x
	replicate	18,703,551	15,863,763 (85%)	2,839,788	18.2x

Table 2.1. PAT-Seq raw sequencing data. Raw reads derived from tissue-specific mRNA libraries on the Illumina Hi-Seq Instrument, mapped to the *C. elegans* WS190 genome annotation.

The overall number of genes and their ratio detected in each biological replicate, with the exception of our negative control, was comparable (~1 vs 0.7) (**Table 2.2**), suggesting that our approach was consistent. In the intestine, we detected a much larger number of genes (7,355 genes) compared with that of pharynx (3,094 genes) and body muscle (3,604 genes) tissues.

samples (tissue)		genes		isoforms	
intestine	*experiment*	7,971	7,355(*)	8,987	8,519(*)
	replicate	8,254		9,432	
pharynx	*experiment*	4,188	3,094(*)	4,427	3,650(*)
	replicate	3,998		4,362	
body muscle	*experiment*	3,404	2,604(*)	3,610	3,024(*)
	replicate	3,478		3,679	
myo-2Δpab-1 (- control)	*experiment*	796	1,011(*)	826	1,120(*)
	replicate	1,146		1,247	

Table 2.2. PAT-Seq mapped data. Mapped reads from the tissue-specific mRNA libraries on the Illumina Hi-Seq instrument. Genes and isoforms are mapped to the *C. elegans* WS190 genome annotation. Genes and isoforms marked with an asterisk correspond to genes and isoforms enriched in both biological duplicates.

Genes and their expression levels between biological replicates correlated well, with the exception of the *myo2p*::Δ*pab-1* control, further supporting the reproducibility of our approach and suggesting that PAT-Seq is specific and sensitive (**Figures 2.4 and 2.5**). A closer look into transcripts recovered with *myo2p*::Δ*pab-1* revealed an enrichment of ncRNAs and other common contaminants, reinforcing that this sample represented random non-specific RNA pulled-down during the immunoprecipitation step (**Figure 2.5**).

We have validated the tissue localization of selected tissue-specific genes identified with PAT-Seq by cloning their promoters and using them to drive expression of GFP *in vivo* (**Figure 2.6 and Table 2.3**).

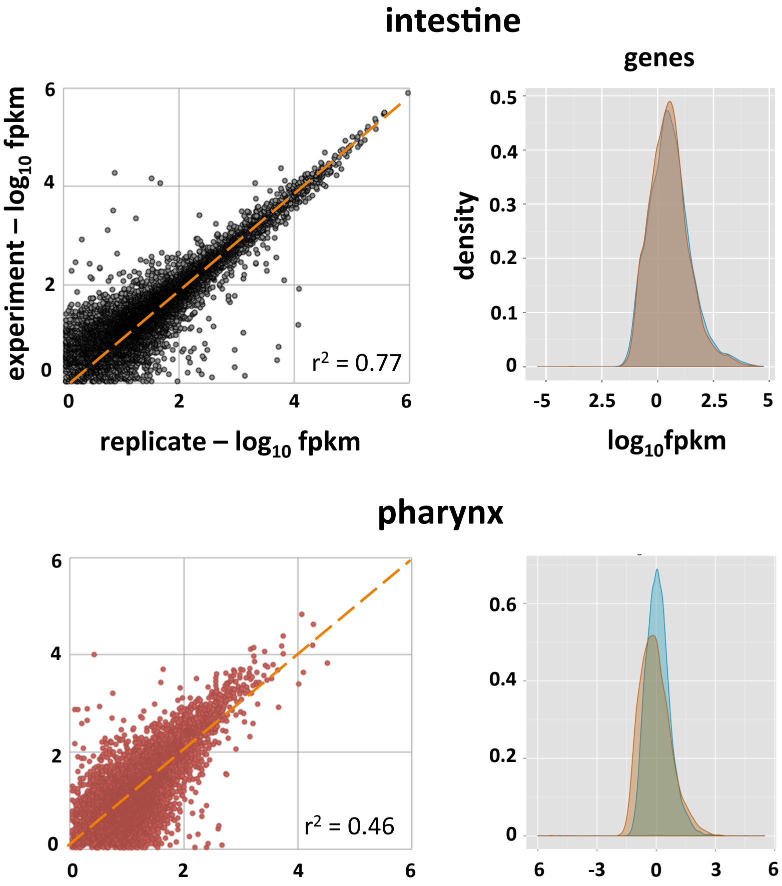
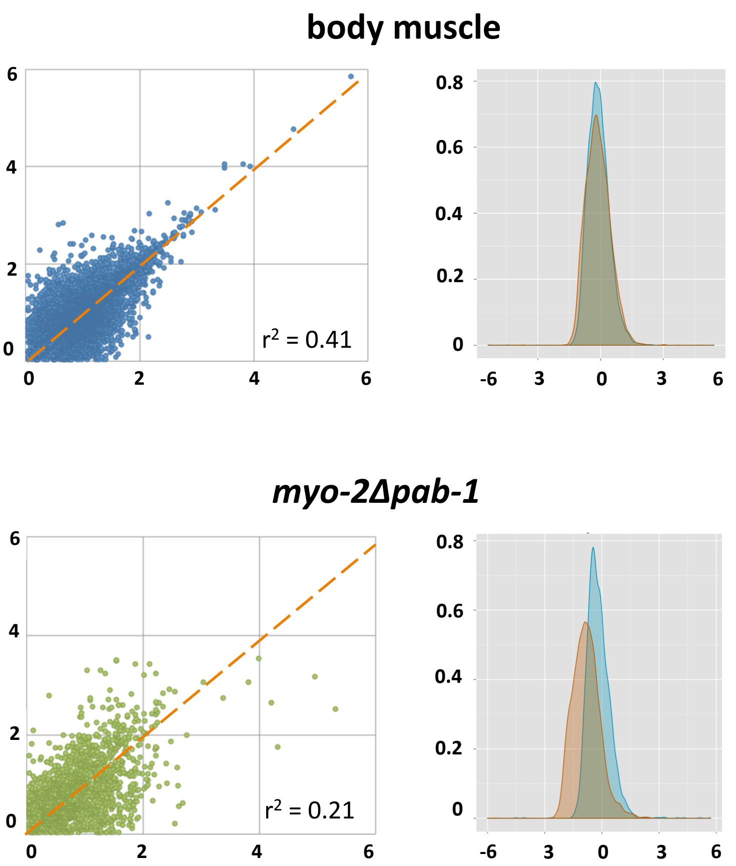


Figure 2.4. Intestine and pharynx PAT-Seq sequencing results. Scatter plot of mapped genes from intestine and pharynx datasets displayed by fpkm value detected in each replicate on a logarithmic (log10) scale to highlight similarity of detection between replicates. The trendline (yellow) displays the expected distribution for 100% similarity between replicates. The right panels show the distribution of the fpkm values in control and replicate samples for each tissue. The plots were generated using the cummeRbund package v. 2.0.


body muscle

myo-2Δpab-1

Figure 2.5. Body muscle and control PAT-Seq sequencing results. Scatter plot of mapped genes from body muscle and *myo-2Δpab-1* datasets displayed by fpkm value detected in each replicate on a logarithmic (log10) scale to highlight similarity of detection between replicates. The trendline (yellow) displays the expected distribution for 100% similarity between replicates. The right panels show the distribution of the fpkm values in control and replicate samples for each tissue. The plots were generated using the cummeRbund package v. 2.0.

We detected a common core set of ~1,500 unique genes present in all three tissues. These transcripts include housekeeping genes such as actin, histone genes, ribosomal proteins, genes involved in transcription, basal transcription factors and genes involved in DNA maintenance and replication. A complete list of genes and isoforms detected in each tissue is shown as diagram in **Figure 2.7, top panel**.

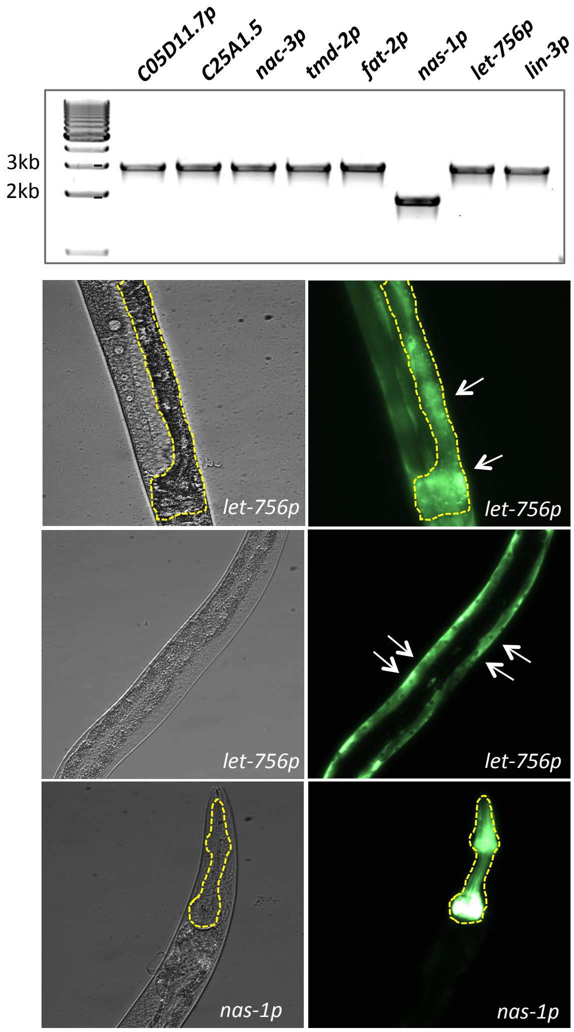
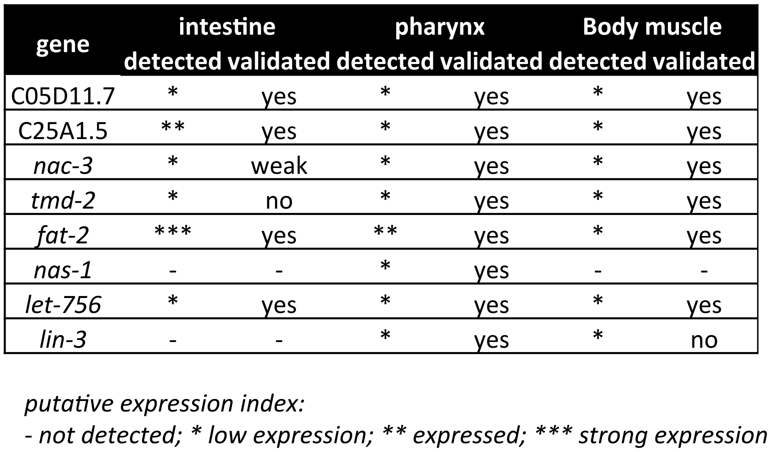


Figure 2.6. Validation of tissue-specific genes detected by PAT-Seq. We have cloned promoter regions for eight tissue-specific genes and used them to drive in vivo expression of our PolyA-Pull plasmid containing GFP. *Top Panel*: Electrophoresis results of PCR confirming the cloning of each of eight promoters upstream of GFP. *Bottom Panel:* Three selected images of transgenic worms expressing GFP in intestine (top panel) and body muscle (middle) driven by *let-756* promoter, and pharynx (bottom) driven by *nas-1* promoter.



gene	intestine		pharynx		Body muscle	
	detected	validated	detected	validated	detected	validated
C05D11.7	*	yes	*	yes	*	yes
C25A1.5	**	yes	*	yes	*	yes
nac-3	*	weak	*	yes	*	yes
tmd-2	*	no	*	yes	*	yes
fat-2	***	yes	**	yes	*	yes
nas-1	-	-	*	yes	-	-
let-756	*	yes	*	yes	*	yes
lin-3	-	-	*	yes	*	no

putative expression index:
*- not detected; * low expression; ** expressed; *** strong expression*

Table 2.3. Table displaying comprehensive results for each of eight promoters driving expression of GFP *in vivo*. The putative expression index reflects the level of expression of each gene obtained from PAT-Seq data in each tissue. We validated gene expression in 19/21 cases using this strategy. Seven out of eight of these genes were detected by our approach in all three tissues. Importantly, the strength of the GFP signal detected in most of the tissues correlate with the expression levels from our sequencing data (data not shown). Out of 21 total experiments, all but two cases were confirmed by GFP expression in the correct tissue (19/21, ~90% of cases). We were unable to detect expression data in the correct tissue for *tmd-2* (intestine) and *lin-3* (body muscle). However these genes may be expressed below the limit of GFP detection. Together, these results provide evidence that PAT-Seq is indeed a sensitive and specific technique to enrich for tissue-specific mRNAs in worms.

We detected a common core set of ~1,500 unique genes present in all three tissues. These transcripts include housekeeping genes such as actin, histone genes, ribosomal proteins, genes involved in transcription, basal transcription factors, and genes involved in DNA maintenance and replication. A complete list of genes and isoforms detected in each tissue is shown as diagram in **Figure 2.7, top panel**.

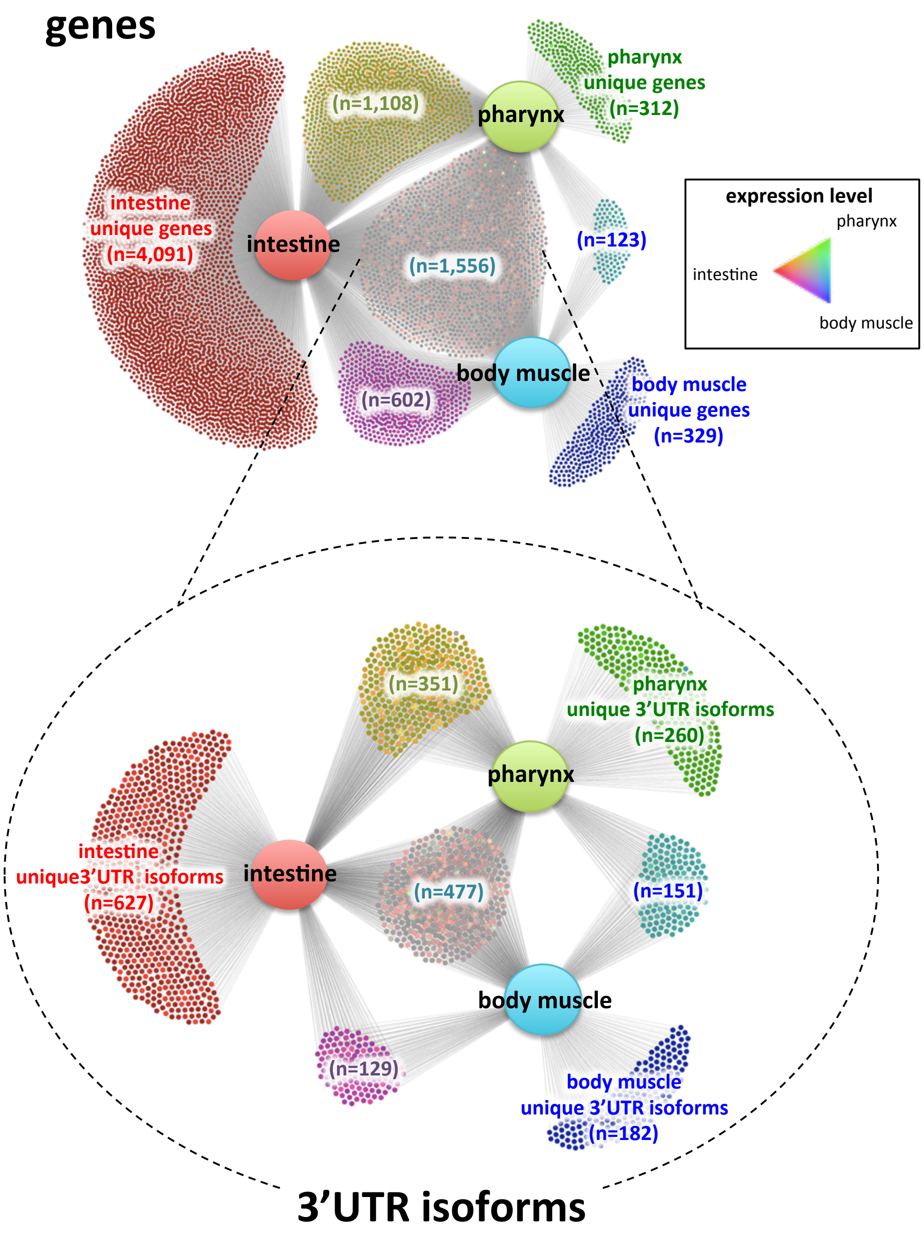


Figure 2.7. Distribution of tissue-specific gene expression and alternative polyadenylation in intestine, pharynx and body muscle. Top panel: Tissue-specific genes identified by PAT-Seq and the distribution of their expression levels between each tissue. A large pool of 4,091 genes is uniquely expressed in the intestine, while a smaller portion of 312 and 329 genes is expressed uniquely in the pharynx and body muscle, respectively. We have detected a common set of 1,556 genes expressed in all three tissues. Edges represent the presence of transcripts in each tissue, and color-coding indicates expression levels of genes in tissues (legend). Bottom panel: The 1,556 genes shared in all three tissues were further sorted based on the 3′UTR isoform and their expression levels. Approximately 25% to 30% of these genes use common 3′ends in these three tissues, while the remaining 70% use tissue-specific 3′UTR isoforms.

The intestine transcriptome is expansive, expressing over 30% of C. elegans mRNAs

The intestine is one of the largest tissues in *C. elegans*, composed of 20 large cells and a total of 30-34 nuclei, with a final 32-fold polyploidy in the adult worm. It is also one of the most functionally diverse tissues, participating in digestion, nutrient transport and storage, innate immunity, response to environmental toxins, defecation, and dauer formation [133-135]. While the intestine transcriptome has been studied extensively [119, 124, 136], our results correlated with and significantly expanded these studies (**Table 2.4 and Figures 2.8, 2.9, 2.10, 2.11, and 2.12**).

datasets	# of genes
WS190	28,122
this study (intestine)	7,361
Haenni *et al.*	3,502
Pauli *et al.*	1,647
McGhee *et al.*	5,623

Table 2.4. Comparative analysis with other available intestine-enriched datasets. We have compared our intestine dataset with Haennei et al., Pauli et al., and McGhee et al. The table displays the total number of genes detected in each dataset.

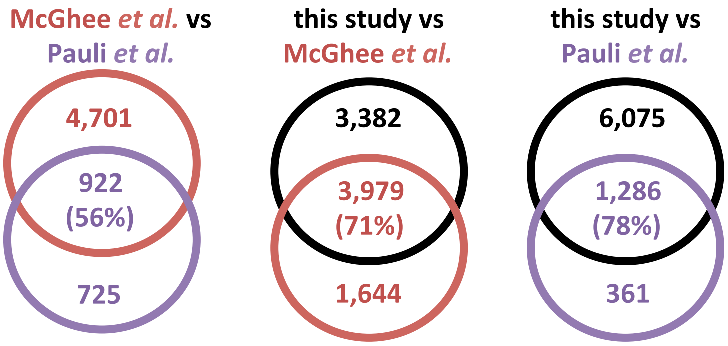


Figure 2.8. The overlaps between our intestine dataset and McGhee and Pauli datasets. Only 56% of genes present in McGhee et al. and Pauli et al. overlap. Our intestine dataset (this study) instead overlaps with these two datasets 71% and 78% respectively.

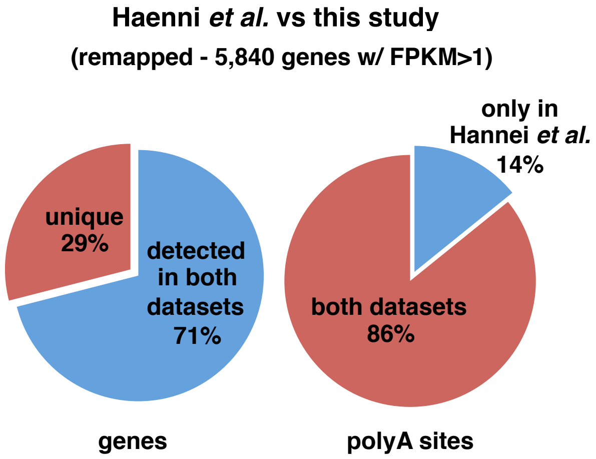


Figure 2.9. Comparison with Haenni et al. datasets. We downloaded and remapped the raw data from the ‘sorted’ dataset produced by Haenni et al., and studied its degree of correlation with our dataset (this study). *Left:* 71% of the top 1,000 genes from the remapped Haenni et al. dataset overlap with intestine genes detected from our study, whereas 29% of genes are only detected in the remapped Haenni et al., dataset. *Right:* 86% (n=6,316) of the polyA sites detected by this study overlap with 3′UTR ends remapped from the Haenni et al. intestine dataset.

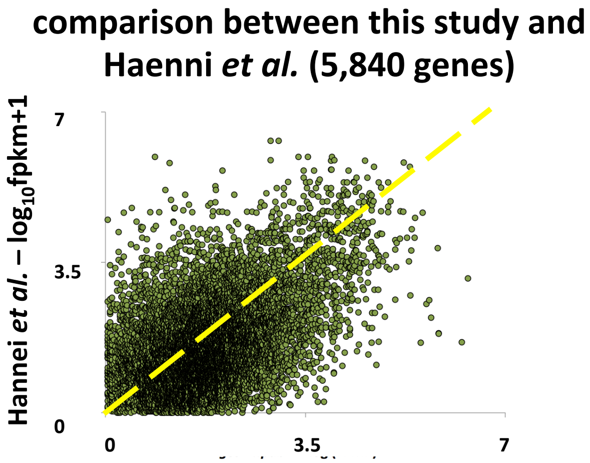


Figure 2.10. Comparison of gene expression levels between genes expressed in both Haenni et al., and our intestine dataset.

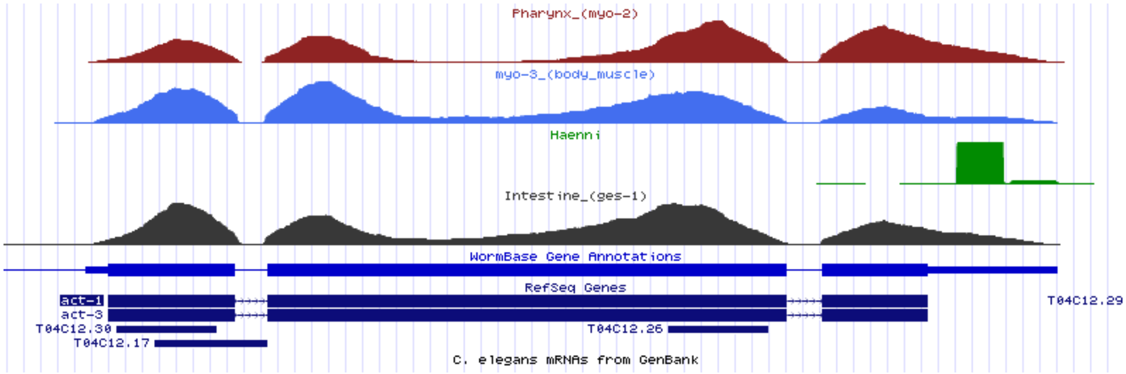


Figure 2.11. Comparison of gene expression coverage. Example of gene expression coverage in our dataset (this study) vs Haenni et al., who used an approach that only mapped the 3′ends of transcripts (red arrow) that are bioinformatically attached to the closest gene model.

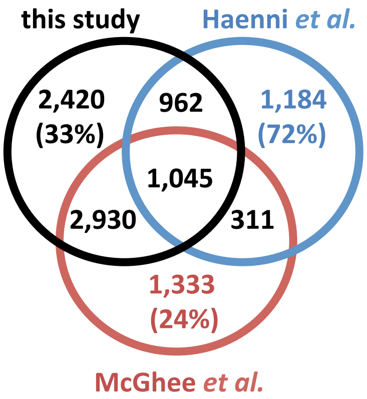


Figure 2.12. Overlap between intestine datasets. The overlap between genes detected in our intestine dataset (this study), Haenni et al. (not remapped), and McGhee et al. datasets. We detected a core set of 1,045 genes that are identified by all three datasets.

We detected a total of 7,355 expressed genes in this tissue (~1/3 of the worm transcriptome), of which 4,091 genes and 4,634 spliced isoforms are uniquely expressed in the intestine, but not in either muscle tissue (**Figure 2.7, top panel**). The most abundant in this dataset were metabolic enzymes and nutrient transport genes, consistent with this tissue's physiological function in digestion. Among these intestine-only genes, we identified 212 unique transcripts that contain a DNA binding domain and were previously described as transcription factors [137, 138]. We speculate that these transcription factors may contribute to the gene regulatory network necessary for tissue identity and function. As expected, members of the GATA family are among the most abundant transcription factors. These factors bind "GATA" elements on promoters and have been shown to regulate endoderm specification and all aspects of intestine development and function in *C. elegans* [136]. In addition, a significant portion (45%) of all transcription factors uniquely expressed in this tissue (96 out of 212) are members of the nuclear hormone receptor family. Many members of this class of transcription factors were also previously shown to regulate *C. elegans* metabolism [139].

We also detected a large pool of novel intestine-specific transcription factors with unknown roles that need to be further investigated.

Recently, Pauli and colleagues coupled mRNA-tagging with DNA microarrays from L4-stage worms and identified 1,647 intestine transcripts [124] (**Figure 2.8**). Others produced a SAGE library of transcripts from dissected worm intestines from mutant adults, and detected a total of 5,623 intestine genes (**Figure 2.8**) [136]. A comparison of each of these datasets with ours identified 71% and 78% overlap with McGhee and Pauli datasets, respectively (**Figure 2.9**), suggesting that the increased sensitivity of PAT-Seq applied to study the intestine transcriptome of mixed stage worms expanded the core *C. elegans* intestine transcriptome.

Haenni and colleagues [119] recently optimized a procedure for extracting intact nuclei from *C. elegans* intestine, followed by fluorescence-activated sorting and deep sequencing of mRNA, focusing on the 3′ end of the transcriptome. This approach allowed the authors to map 3,502 genes expressed in this tissue (**Figure 2.8**). We compared our intestine datasets with these results, downloading and remapping the raw reads from Haenni *et al.* using the same filtering criteria used in our datasets (see Experimental). We found that 71% of genes were detected in both datasets, leaving 29% of genes present in the Haenni *et al.* dataset not present in ours (**Figure 2.9**). Despite these differences, the distribution of genes detected in both datasets plotted by expression values correlated (**Figure 2.10**). Our inability to detect 1/3 of genes in Haenni *et al.* may be attributed to the fundamentally different techniques used in the preparation of our cytoplasmic, intestine-specific mRNAs using Pat-Seq versus nuclear cDNA prepared from isolated FACS sorted nuclei as in Haenni *et al.* (**Figure 2.11**).

The difference in gene pools may have arisen because of the different cellular origin of the mRNAs analyzed by the two studies.

Finally, we have overlapped our datasets with McGhee *et al.* and Haenni *et al.*, the two intestine transcriptomes datasets obtained by sequencing, revealing a core set of shared 1,045 genes (**Figure 2.12**). These 1,045 genes represent a collection of high confidence intestine expressed genes, supported by at least three independent approaches, and will provide important insights in unraveling the genetic basis of intestine tissue identity in worms.

Muscle transcriptomes are smaller but contain mostly unique gene pools

C. elegans possess only two large muscles: the pharynx and the body muscle. The pharynx, or foregut, is an important developmental model composed of eight layers of muscle, in addition to surrounding neural and epithelial tissue [106]. The muscle component of this tissue is composed of 20 cells that coordinate intake and physical crushing of the worm bacterial diet [140], subsequently facilitating raw nutrient transfer to the intestinal lumen for digestion. While the genetic factors controlling early development of the pharynx have been described in detail [141], individual cell-types of the pharynx are less characterized because genetic signatures belonging to these specific subgroups have not been extensively studied.

We detected 3,099 genes expressed in pharynx muscle (**Figure 2.7, top panel**). Among the top genes expressed were several myosin and actin isoforms, and pharynx-specific neurotransmitters, consistent with this tissue's muscular identity. Importantly, we found only 312 unique genes with 338 spliced isoforms significantly expressed in this tissue (~10% of the total dataset) (**Figure 2.7, top panel**). Most of these genes have unknown function, with only 70 transcripts (22%) described in Wormbase so far [142]. The top genes of this list are collagen isoforms and motor protein genes. Within this pool we also detected 13 pharynx-specific transcription factors. Most of their gene targets are unknown.

The body muscle tissue, defined as all non-pharyngeal muscle cells, is homologous to vertebrate skeletal muscle [143], and critical for locomotion, egg laying, defecation, and mating [144]. Several groups also studied the transcriptome of this tissue [121, 122]. We detected 2,610 genes expressed in the body muscle. Similar to our pharynx dataset, within this pool we detected only 329 unique genes corresponding to 365 spliced isoforms (**Figure 2.7, top panel**). The list of top ten genes in the body wall muscle dataset was also enriched for muscle-specific genes such as myosin and actin isoforms. We also detected a unique gene pool, including previously identified genes, such as those in the calveolin family [145], and type IV collagen [146]. Out of predicted worm transcription factors, we identified 22 genes that are uniquely expressed in this tissue, including *unc-120*, a SRF-like transcription factor essential for body wall muscle development, and *blmp-1*, a PRDM family member required for embryonic slow muscle fiber formation in vertebrates [143, 147].

Using PAT-Seq we were able to detect many statistically significant isoform expression changes between intestine and muscle tissues (**Figure 2.13**).

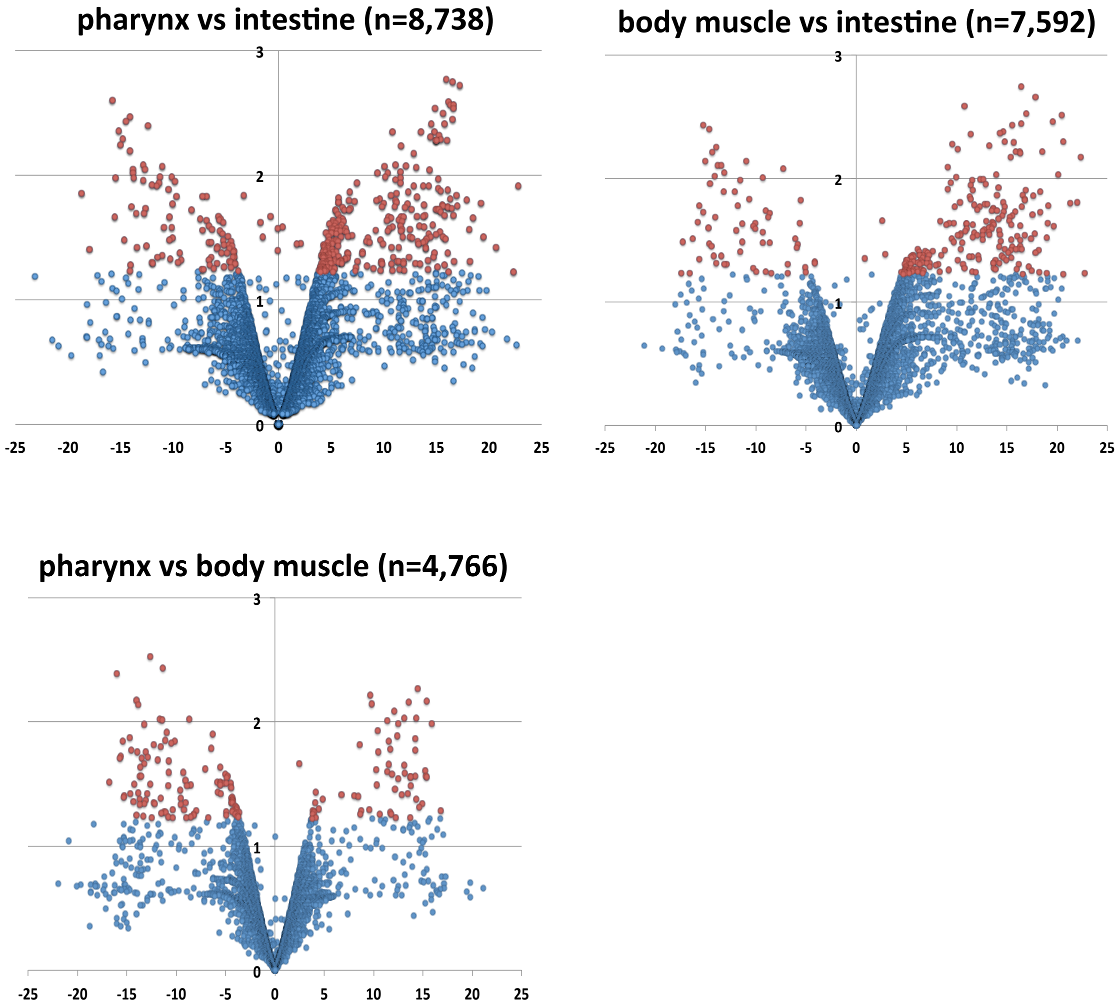


Figure 2.13. Differential mRNA isoform expression analysis. We have studied the changes in mRNA splice isoform expression for genes detected among each combination of two tissues in our datasets. Volcano plots showing the changes of isoform expression between each tissue (p-value versus fold-change). Total number of isoforms that significantly switch between two tissues (p<0.05) are shown in red and the total number of genes in this category are boxed.

Tissue-specific promoters possess unique signatures

The compact nature of the promoter regions in the *C. elegans* genome provides us with a unique platform to examine tissue-specific elements in these regions and highlight signatures such as transcription factor binding sites. We first studied the sequence composition of these promoters, defined as the portion of genomic sequences from -500 to +100 from the TSS of protein coding genes [119]. We compared these sequences to promoters of random genes from the whole *C. elegans* transcriptome (28,122 genes from WS190) (**Figure 2.14**). We found that, contrary to higher metazoans, such as in humans where promoters are significantly enriched in cytosines and guanosines, *C. elegans* promoters are significantly enriched in adenosine and thymidine. These two nucleotides represent more than 66% of the total nucleotide composition in these promoter regions (*data not shown*). We also detected a strong T-rich region closer to the transcription start site of intestine genes, which perhaps is implicated in intestine-specific mechanisms of transcription initiation (**Figure 2.14**).

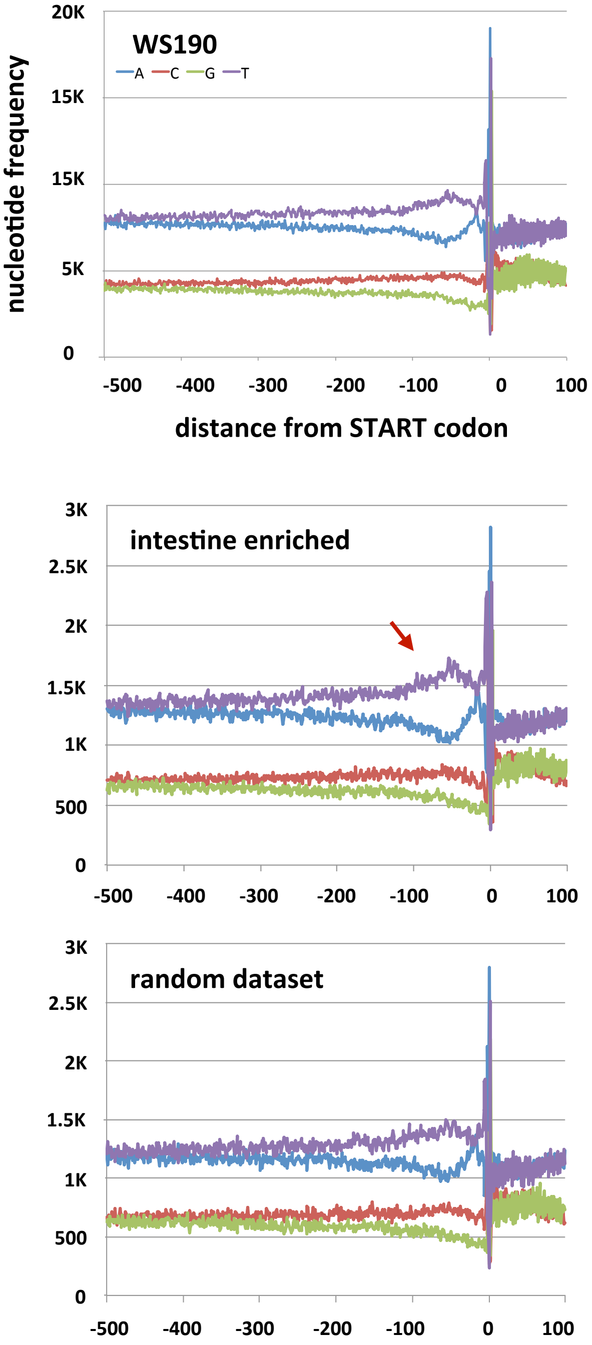


Figure 2.14. Nucleotide enrichment in of promoter regions for intestine expressed genes. We extracted and studied the DNA regions 500bp upstream from the start codon for each of 4,095 genes unique in our intestine dataset. The average base composition for promoter regions in all WS190 transcripts (top), intestine-specific transcripts (middle), and a random dataset of 4,095 genes (bottom). We detected a strong enrichment of thymidine within 100nts upstream of the transcription start site (red arrow).

We then scanned these promoter regions for enriched elements uniquely present in this tissue, hoping to detect tissue-specific signatures. We calculated the frequency of all possible hexamers within the promoter regions of our intestine and random datasets, and then gathered the frequency of these elements into six bins consisting of 100 nucleotides each (**Figure 2.15**). Among the top hits, we detected a significant enrichment of many 'GATA' binding sites in these promoters (24% to 40% higher) (**Figures 2.15 and Table 2.5**).

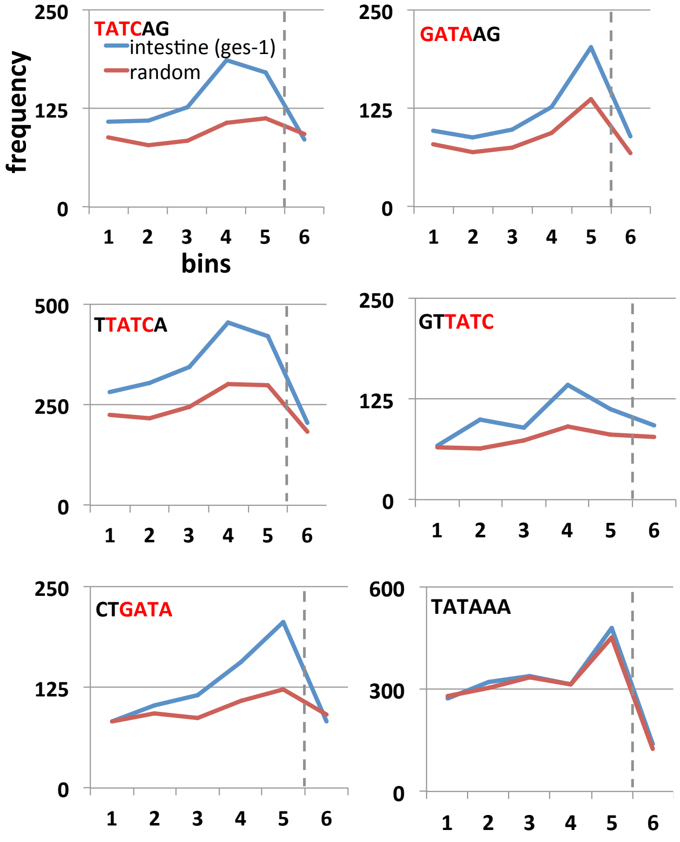


Figure 2.15. Enrichment of hexamers in promoters of intestine expressed genes. The conserved ‘GATA’ or its antisense ‘CTAT’ element in this tissue compared with the same number of randomly selected promoters. The canonical ‘TATAAA’ (TATA-box) was used as a comparison (bottom right) to show equal enrichment of this hexamer in both intestine and random sets.

motif	random dataset	intestine enriched (% increase)
TATCAG	563	787 (+40%)
TTATCA	1,469	2,011 (+37%)
GTTATC	451	602 (+35%)
GATAAG	524	701 (+33%)
CTGATA	583	745 (+28%)
GATAAC	409	519 (+27%)
TGATAA	1,291	1,599 (+24%)

Table 2.5. List of hexamers with the conserved 'GATA' element in our intestine dataset. The percent enrichment of this element over the set of randomly selected genes is shown. ***P<0.001, by 2-tailed Chi-square test.

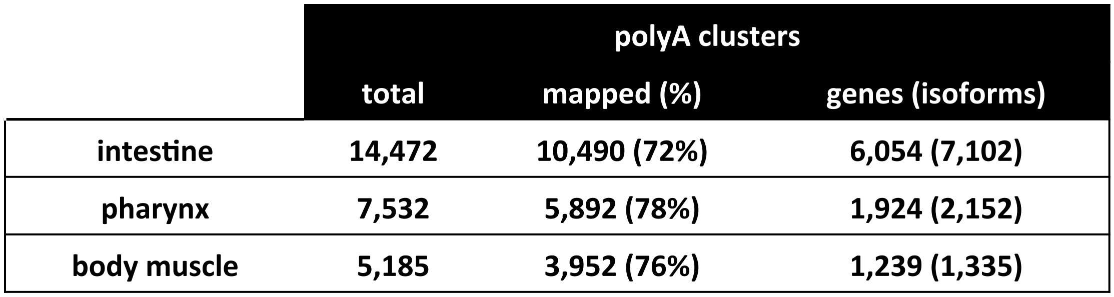
We expanded these analyses in the muscle tissues and scanned promoter regions for enriched hexamers and known *trans*-acting factors (*data not shown*). Though we used unique gene datasets to search for tissue-specific motifs, the body muscle and pharynx shared several highly significant sequences, suggesting the existence of common core regulatory elements modified for pharynx and body muscle (*data not shown*).

Many eukaryotic promoters contain a 'TATAA' binding element used to recruit the transcription machinery to the transcription start site. When we extended this search in promoter regions in the worm genome (WS190) and in our three tissues, the frequency of the 'TATAA' box was ~37%, slightly higher than what is observed in human (24%) [148] (*data not shown*), suggesting that while not ubiquitous, the TATA box is still abundant in nematodes.

3′end formation in intestine and muscle tissues is unlikely driven by tissue-specific sequences

We next studied changes in polyadenylation signal elements (PAS) and APA in these datasets. APA is pervasive in *C. elegans* [88, 89], but it is still unclear in which tissues these 3′UTR isoforms are expressed, how they are produced, and their consequences. We employed an innovative library preparation method based on isothermal linear amplification of polyA+ RNA, which allowed us to bypass ligation-based approaches and precisely detect both the transcriptome and 3′UTRome of selected tissues profiled at the same time (see Experimental). This method, named SPIA, produces continuous linear synthesis of ssDNA amplicons from a single RNA template, producing consistent read numbers through the transcriptome and minimizing internal mis-priming that could generate false 3′ends during the cDNA library preparation [132].

Using this approach, we were able to build ~20,000 high-quality PAS clusters (72% to 78% of the total mapped PolyA clusters) that allowed us to map 3′UTR ends at single base resolution for ~6,000, ~2,000, and ~1,200 genes in our intestine and two muscle datasets, respectively (**Table 2.6**). Importantly, more than 80% of these mapped 3′UTR isoforms overlapped with previously described datasets [88, 89] (**Figure 2.17**), strongly suggesting that the vast majority of 3′UTRs detected using our approach are *bona fide* 3′ends of mRNAs.



	polyA clusters		
	total	mapped (%)	genes (isoforms)
intestine	14,472	10,490 (72%)	6,054 (7,102)
pharynx	7,532	5,892 (78%)	1,924 (2,152)
body muscle	5,185	3,952 (76%)	1,239 (1,335)

Table 2.6. Summary of 3′UTR poly-A site mapping in tissue datasets. We used the raw sequencing reads to map high-quality polyA sites onto the WS190 worm annotation and compared our results with two published *C. elegans* 3′UTRome datasets. The number of polyA clusters mapped from polyA-containing sequencing reads (total), the portion of those that mapped to the WS190 worm genome annotation (mapped), the number of genes with polyA sites mapped (closest gene to the polyA cluster) and the number of isoforms resulting from distinct mapping of polyA clusters (isoforms).

We then studied the length of the 3′UTRs in these tissues and found that, on average, intestine genes possess shorter 3′UTRs, when compared to pharynx and body muscle genes (**Figure 2.17**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

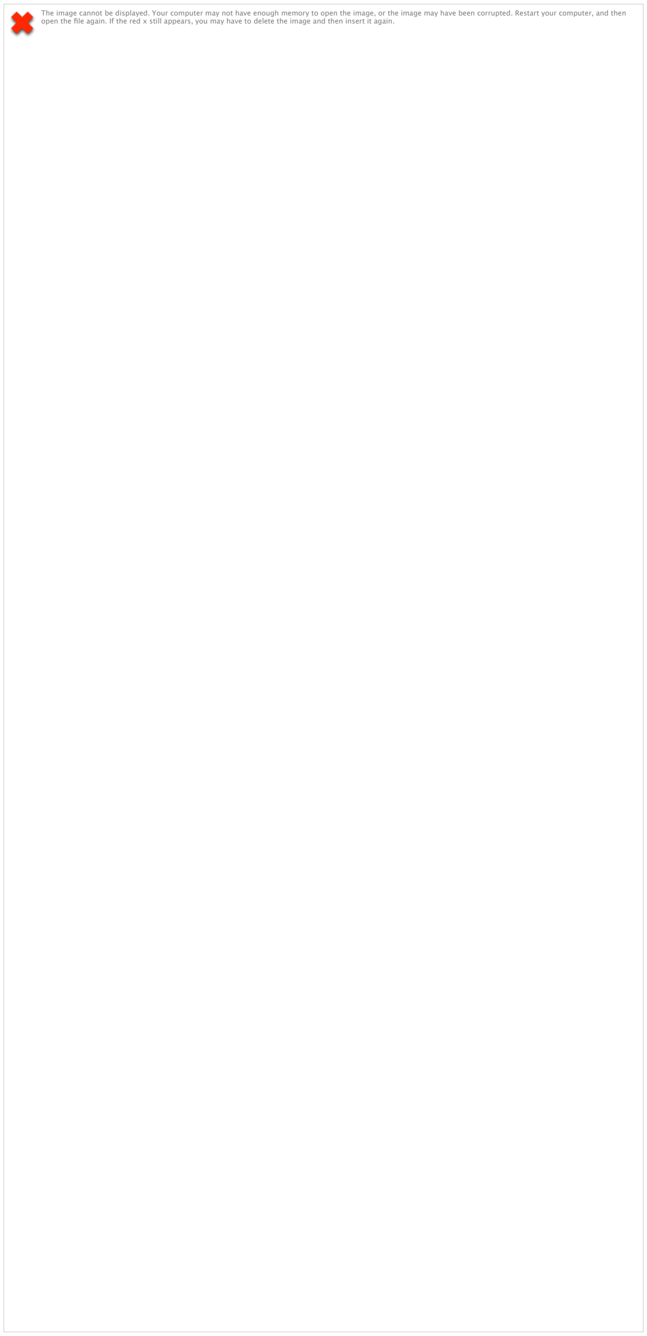
Figure 2.16. Poly(A)-cluster comparison and 3′UTR length analysis. The majority of mapped 3′UTR isoforms are supported by two published 3′UTRomes and almost 90% of them are supported by at least one dataset. Left panel: the percentage of isoforms mapped to either of two published 3′UTRomes (green and red), to both (blue), and those not present in either 3′UTRome dataset (purple). Right panel: the distribution of 3′UTR length for all 3′UTR isoforms found in each tissue dataset, along with the median (vertical dashed red line) and the average length.

In mammals, shorter 3′UTRs tend to escape post-transcriptional gene regulation and are more stable in comparison with longer mRNAs [149]. This activity has not yet been documented in worms, but presumably these short 3′UTRs could lead to an increase in protein translation in the *C. elegans* intestine to support its diverse physiological roles.

Next, we analyzed the PAS elements, which are sequences in 3′UTRs known to direct 3′end formation. We superimposed our tissue-specific datasets to the worm 3′UTRome from the modENCODE project [88] and extracted the PAS nucleotide composition (**Figure 2.17**). The canonical PAS element 'AAUAAA' in intestine and muscle tissues was ~10% more abundant than in the 3′UTRome overall (**Figure 2.17**). The PAS sequences containing one permutation of the canonical element were similar in all three tissues, while those containing two or more permutations were drastically reduced (**Figure 2.17**). PAS position within 3′ ends of mRNAs was similar in all three tissues (**Figure 2.18**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 2.17. Analysis of PAS usage between tissues. We extracted the PAS elements present in the 3′UTRome and assigned them to 3′UTR isoforms present in each of our tissue datasets. Each chart represents the percentage of distinct isoforms present in each tissue dataset containing the canonical PAS ‘AAUAAA’ (blue), seven of the next most common PAS elements, and the remaining ‘other PAS’ sites. The nucleotide changes from the canonical PAS element are highlighted in red.

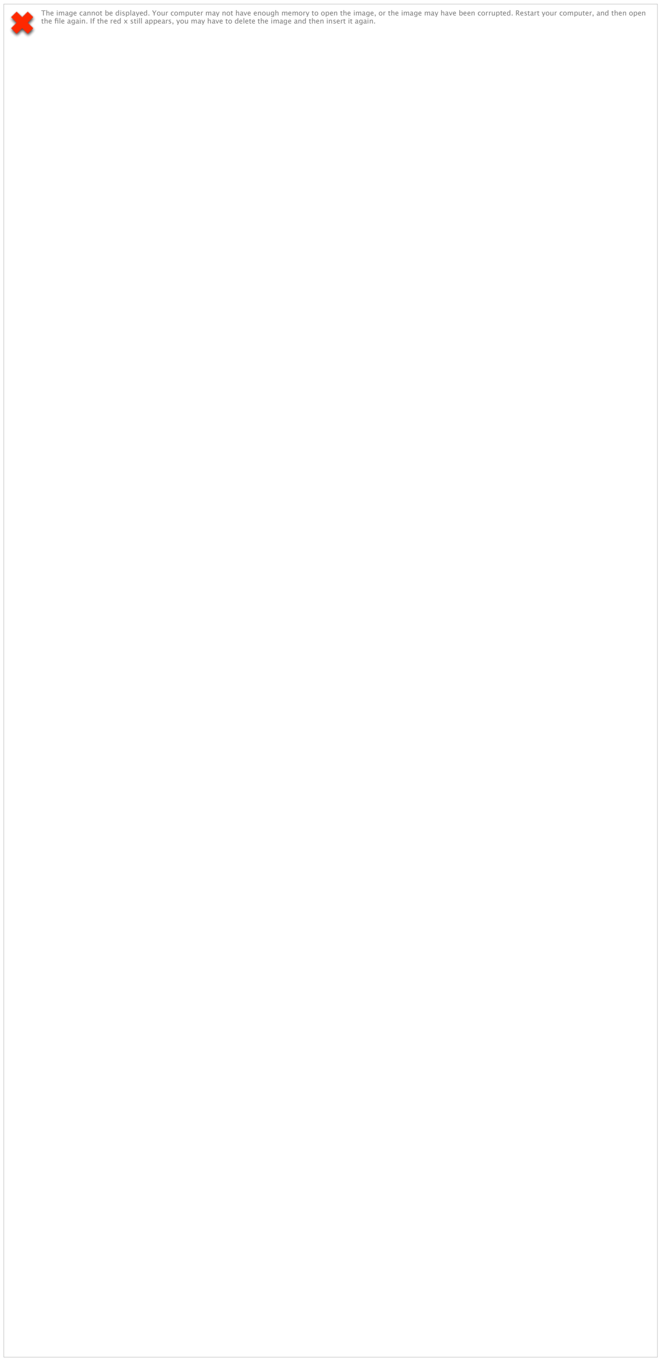


The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 2.18. PAS location in reference to the cleavage site. Plot showing the position of the canonical PAS 'AAUAAA' and other PAS sites with reference to the cleavage site for all isoforms in each tissue for 3′UTRs in common with Mangone et al. All three tissues have a strict positional requirement for PAS elements at -19nts from the cleavage site.

We then studied the sequence conservation near the mRNA cleavage sites in genes present in each dataset, hoping to detect tissue-specific signatures (**Figure 2.19**). The nucleotide frequency in these regions was remarkably similar between tissues (**Figure 2.19**). We detected only a slight change in frequency of adenosines near the PAS site, which was specific to 3′UTRs expressed in the intestine (**dashed box in Figure 2.19**).

Taken together, our results suggest that these sequences may not contain elements important in tissue-specific 3′ end formation, or that such elements are further downstream of the cleavage site and not detected by our analysis.

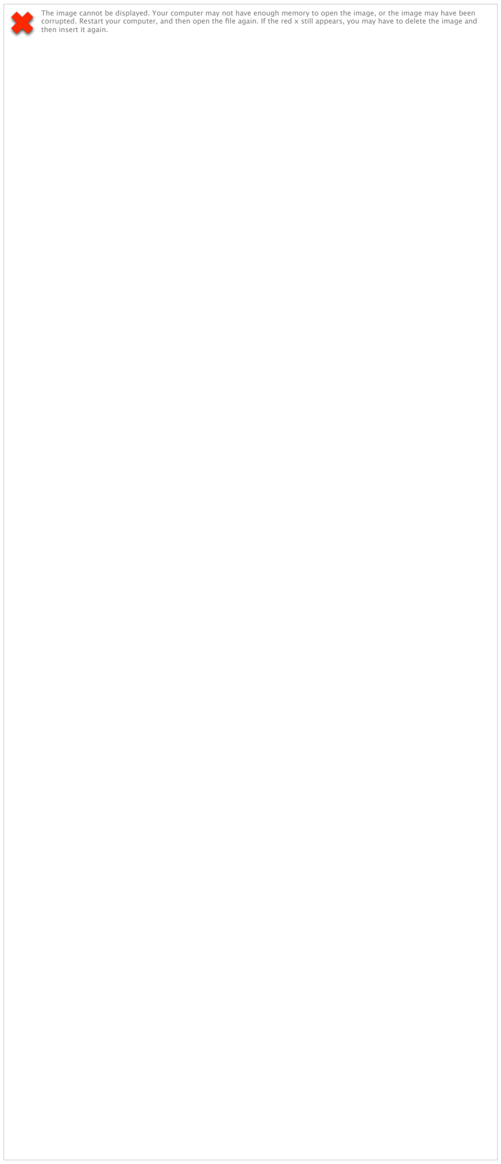


The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 2.19. Nucleotide frequency distribution within the cleavage sites. While all these tissues have a very similar pattern, intestine genes have an unusual larger adenosine-enriched block at -20nts from the cleavage site (dashed box).

Alternative polyadenylation is pervasive in intestine and muscle tissues

The two available worm 3′UTRome datasets estimate that ~46% of *C. elegans* genes use APA [88, 89]. APA is coordinated through development, where proximal 3′UTRs are expressed in earlier developmental stages and distal are expressed more frequently in later developmental stages [88]. However, the extent to which APA is coordinated between *C. elegans* tissues and how it may participate to establish cell identity has not yet been addressed. We employed a normalization method to select for higher confidence 3′UTR isoform switching events based on the ratio between PAS coverage (see Experimental). Intestine tissue has a larger pool of genes with two or more 3′UTR isoforms (twice as many genes as in muscle tissues), while muscle tissues mostly use single 3′UTR isoforms (**Figures 2.20 and 2.21**).



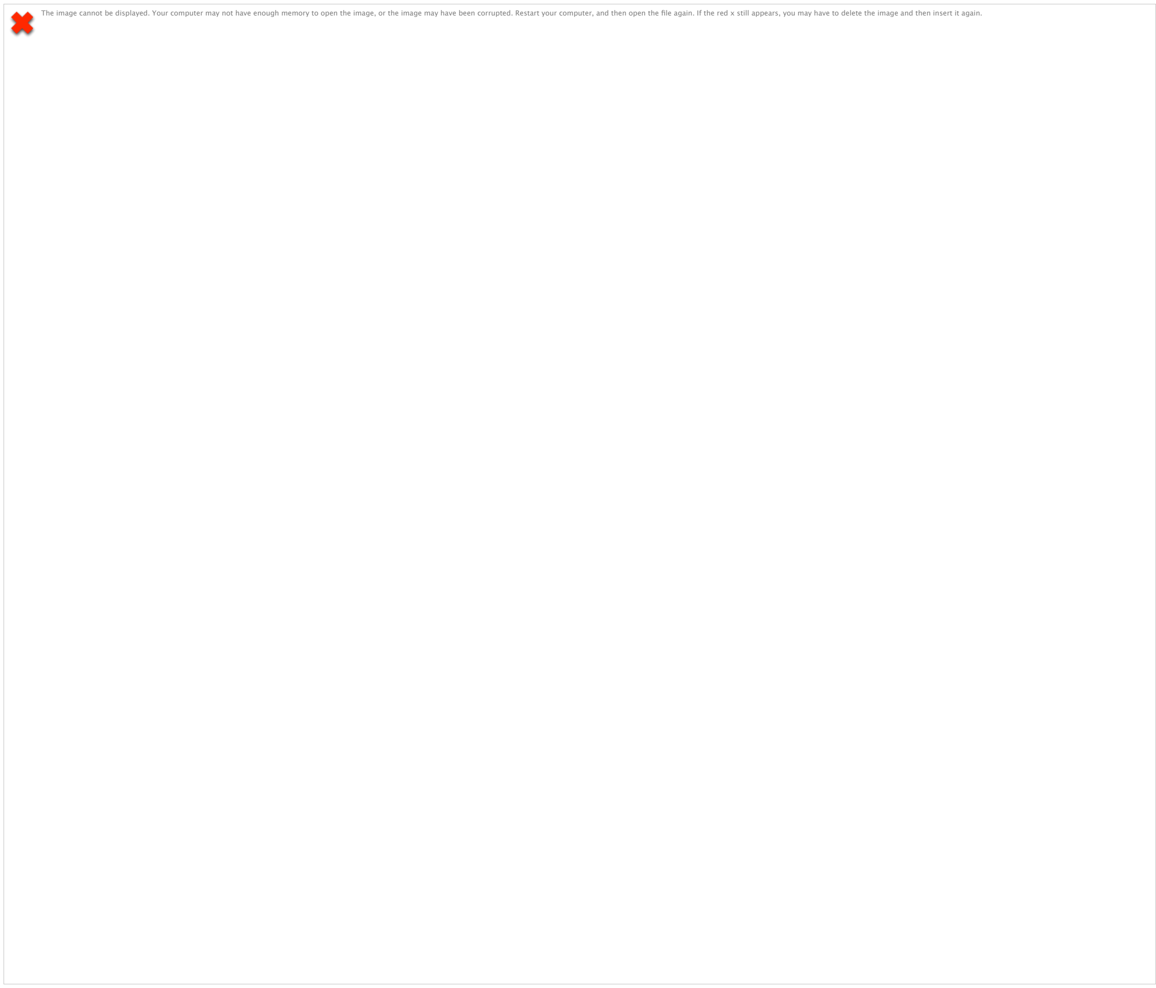
The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 2.20. Abundance of APA in *C. elegans* tissues. A finalized list of genes with mapped 3′UTR isoforms was generated for each tissue and used to compare the abundance of 3′UTR isoforms between tissues. Proportion of genes subject to alternative polyadenylation in each tissue. The intestine expressed significantly more genes containing more than one 3′UTR isoform, while the muscle tissues expressed similar proportions of genes with more than one 3′UTR isoform.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 2.21. Isoforms per gene in *C. elegans* tissues. The average number of 3′UTR isoforms detected for each gene/tissue. The number of genes and isoforms (frequency) are displayed in each column (left x-axis). We calculated and displayed the change in 3′UTR isoform to gene ratio (right x-axis) between each tissue (green trend line). We detected slightly more APA in the intestine and pharynx, when compared with the body muscle tissue.

We reasoned that if APA is a tissue-specific event, we would be able to detect it by following the dynamics of 3′end formation in genes with one 3′UTR isoform detected in each tissue. Indeed, we found that 18-26% of those genes switched 3′UTR isoform in a tissue-specific manner (**Figure 2.22**). Interestingly, intestine genes more often used distal PAS sites, while both muscle tissues used proximal PAS sites. When we focused this analysis comparing 3′UTR switches in genes expressed in all three tissues, we found that ~25% of these genes use APA in a tissue-specific manner (**Figure 2.22**), suggesting that tissue-specific APA in worms is abundant.

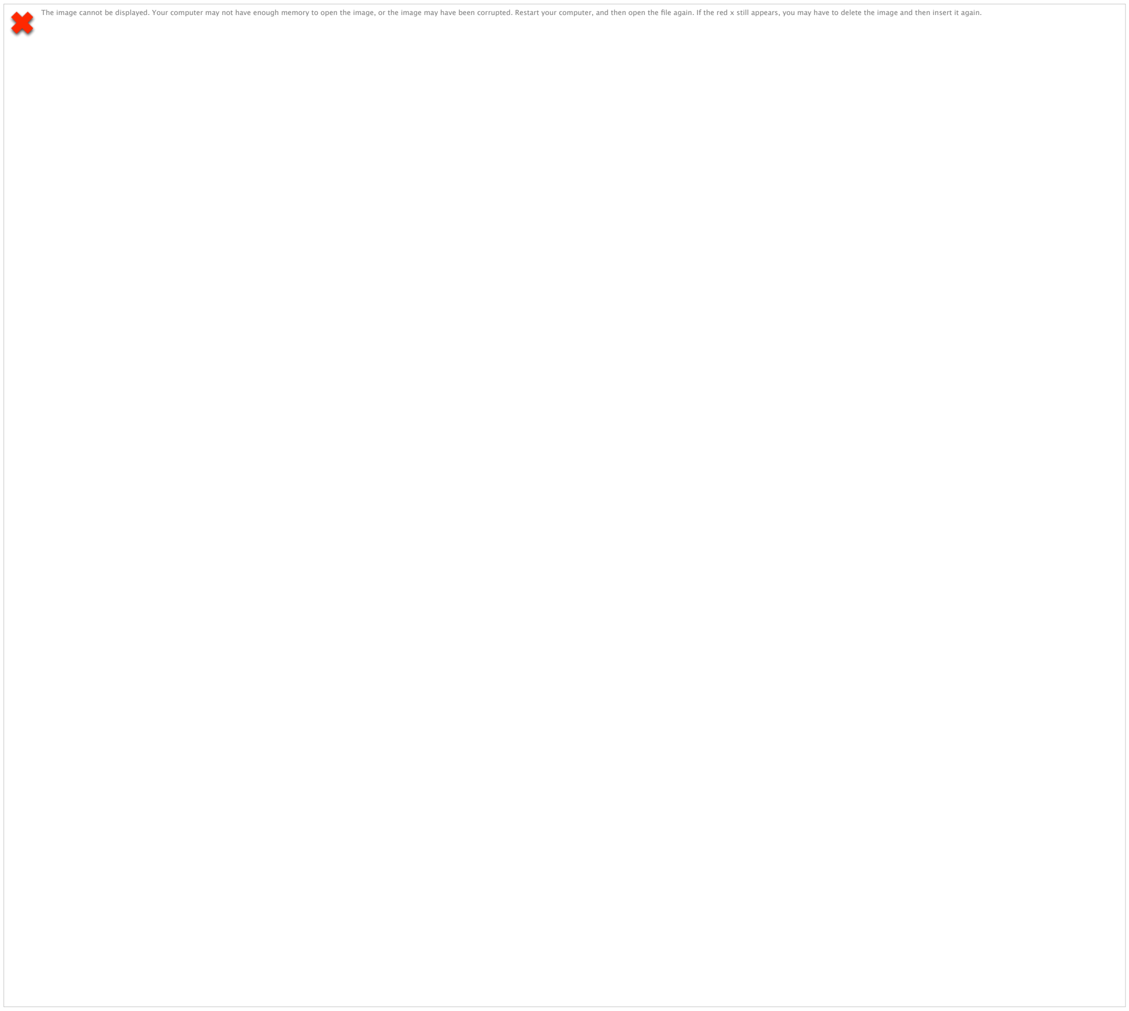


The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 2.22. APA is pervasive between *C. elegans* tissues. We have followed 3′UTR length changes in genes with only one 3′UTR isoform between intestine, pharynx and body muscle tissues. Length comparison between the same genes expressed between intestine and pharynx, pharynx and body muscle and intestine and body muscle tissues. Shaded circles represent those genes expressed with proximal 3′UTR isoforms in the intestine (black), pharynx (red) or body muscle (blue), where the distal isoform was detected in the other corresponding tissue in each graph. Genes with 3′UTR isoforms that were the same length between each tissue are represented in grey as noted in the legend. *Lower right:* Distribution of unique 3′UTR isoforms for genes detected in all three tissues. The majority of these 3′UTRs are common in all three tissues (blue). Genes with a 3′UTR isoform in the intestine distinct from muscle tissues are also abundant (muscle shared). Only 2% of these genes express different 3′UTR isoforms between all three tissues (distinct).

We searched within our three datasets for commonly expressed genes with tissue-specific 3′UTR isoforms (**Figure 2.23, right**).

While muscle tissues had a similar set of genes with tissue specific 3′UTR isoforms, the intestine tissue had, on average, twice the amount (**Figure 2.23, right**), suggesting that the *C. elegans* intestine uses more APA than the two muscle tissues.



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 2.23. Analysis of tissue-specific 3′UTR isoforms. We calculated the proportions of genes in each tissue that have tissue-specific 3′UTR isoforms and how many of these 3′UTRs have predicted microRNA targets. *Left:* Charts displaying the proportion of genes containing tissue-specific 3′UTR isoforms (blue). The intestine expresses approximately two times as many tissue-specific 3′UTR isoforms as muscle tissues. *Right:* We compared the proportion of microRNA targeted genes with tissue-specific 3′UTRs (blue) to the same number of randomly selected genes (grey) in each tissue. Significantly more genes with tissue-specific 3′UTR isoforms have microRNA targets. microRNA targets were predicted using PicTar Software, using three species and five species conservation criteria, and from ALG-1 pull-down experiments (Zisoulis et al. 2010 [32]). *P-value <0.05, **P < 0.01, based on two-tailed Student’s t-test.

Since we were able to detect widespread tissue-specific APA, we were interested to study if the genes that use tissue-specific APA were enriched with miRNAs targets, and perhaps use APA to escape their regulation. We searched in each tissue for genes with tissue-specific 3′UTR isoforms that have bioinformatically predicted microRNA targets (**Figure 2.23, right**). Remarkably, genes with tissue-specific 3′UTR isoforms were enriched with miRNA targets using both a 3-species (~49%) and a 5-species (~25%) conservation filter (**Figure 2.23, right**). This is significantly more abundant than the average number of total genes expressed in each tissue having miRNA targets using the same criteria (**Figure 2.23, right**).

miRNA prediction software produces significantly high false and negative hits that cannot be used to properly assign targets [53]. When we instead compared our dataset to *in vivo* miRNA target footprint data from past studies [127], we found a similar enrichment, with an average of 21% of genes with tissue-specific 3′UTR isoforms containing validated targets (**Figure 2.24**).

In conclusion, these data suggest that miRNA targets are much more abundant in ubiquitously expressed genes with tissue-specific 3′UTR isoforms than in genes without APA, strongly linking APA to miRNA targeting and post-transcriptional gene regulation. Our data supports a model whereby the 3′UTRs of genes with APA are regulated in a tissue-specific manner in order to evade or participate in microRNA targeting.

APAome.org, a tissue transcriptome resource for C. elegans biology

We have made our data publically accessible through our APAome website [128]. The APAome includes our tissue-specific datasets, as well as other important worm 3′UTR datasets [88, 89, 119], allowing the community to have a comprehensive view of APA and 3′UTR biology in worms. The APAome database provides detailed information on 3′UTR isoforms for all protein-coding mRNAs present in Wormbase [142], novel PicTar [88] and TargetScan [89] predictions, and includes annotations extracted from other databases as well as new annotations generated by others.

Discussion

In this study, we have coupled mRNA-tagging with high-throughput sequencing in a novel technique that we called PAT-Seq, and used it to perform an integrative analysis of the mRNA transcriptome of *C. elegans* intestine, pharynx, and body wall muscle tissues. We have studied their transcriptome and 3′UTRome at an unprecedented resolution. In addition, since these three libraries were prepared using the same approach, we were able to directly compare the changes in gene expression and the gene content across these three somatic tissues, without extrapolating data from other studies. Our approach is an improvement over past methodologies [129, 130], allowing the identification of tissue specific mRNAs at higher resolution.

PAT-Seq highlights C. elegans intestine and muscle transcriptome dynamics

We found that the intestine transcriptome is significantly larger than in muscle tissues, possibly to support its especially diverse physiological roles. Intestinal cells are much larger and are increasingly polyploid throughout larval development, with more transcriptional capacity compared to the smaller, diploid muscle cells [150]. Although there are no other comparative data in worms, recent genome-wide transcriptome analyses in the human intestine track support our findings, showing that more than 75% of all protein-coding genes are expressed in this tissue [151]. The twenty large, intestinal cells may require a large pool of distinct genes to carry out functions specific to their anatomical location, since altogether these cells span from the pharynx to the posterior of the animal and the intestine is one of the largest tissues in worms.

Overall, intestinal tissue is more different in gene composition than the two profiled muscle tissues. In intestine, the most abundant genes detected are common metabolic enzymes, such as *fat-1*, *pmt-1*, *asp-1,* and others, which were also detected in other available intestine-enriched datasets [119]. Importantly, genes and isoforms detected in our intestine dataset correlated with, and significantly expanded, previously described worm tissue-specific datasets [119, 124, 136]. In the pharynx and body muscle, the top genes detected were myosin genes, actin isoforms, and other genes. We also detected a large number of tissue-specific alternative splicing events and fold change differences in gene expression for genes in common between tissues.

Gene expression changes could be caused by stage-specific enrichments in our pull-down experiments. We have prepared our tissue specific RNAs using a well established protocol that allows the growth of worms in liquid media [152] (www.wormbook.org).

This protocol is known to provide an even representation of each worm developmental stage. Since all our samples were prepared using the same protocol, it is unlikely that a given sample is biased towards a specific developmental stage. Importantly, our intestine dataset overlaps consistently with recently published studies [119, 124, 136], suggesting that if there is a general bias in all three tissue-specific datasets, it is very low.

Our study detected many tissue specific genes in intestine and muscles that were previously reported by others in the same tissues (Wormbase). We have also validated a selected portion of these hits using a GFP reporter approach (**Figure 2.6**). We think that although very sensitive and reproducible, our PAT-Seq approach may introduce some noise at the lower end of detection. Further experiments may need to be performed to validate the tissue specific localization of these low expressed genes.

C. elegans promoters are AT-rich and contain tissue-specific motifs

Our promoter analysis showed that worm promoters are AT-rich. This result is consistent with what others have found in worm genomic regions [153], Drosophila [154], and Xenopus [155], and very different from what is observed in mammalian promoters, which are GC-rich [156]. GC-rich regions in promoters increase genomic thermostability [157], provide more binding motifs for transcriptional activators [155], and support promoter gene silencing through DNA methylation at GC islands. The AT-rich nature of *C. elegans* promoters was previously observed in *C. elegans* genes expressed in the germline [158], and perhaps reflects a simpler model of transcription initiation with reduced regulation.

We also demonstrate that this approach can effectively identify previously reported and novel sequence elements in tissue-specific gene datasets. As a *proof-of-concept*, we detected GATA transcription factors known to be critical for all aspects of *C. elegans* intestine development and adult function in our intestine transcriptome [119, 124, 159], and detected potential GATA sites present at a higher frequency in the promoters of intestine-specific genes. Importantly, our study highlighted the presence of many novel enriched sequence motifs, many of which have not been described yet in the literature. While we were able to predict several transcription factors that could recognize these motifs, we still do not know if they are indeed functional, and further *in vivo* studies need to be performed to further characterize their role.

SPIA library preparation increases yield and robustness of polyA sequencing

We have sequenced our tissue-specific libraries using a proprietary library preparation method, named Single Primer Isothermal Amplification (SPIA), which is ideal for use with mRNA-tagging, since the RNA yield from this approach is typically low [132, 160]. Unlike recently developed methods used to map 3′UTRs, such as 3P-Seq [89] and FANS/3′-end-seq [119], SPIA generates cDNA libraries that cover the entire transcript, allowing for more extensive downstream transcriptome analysis within the same experiment, such as coupling gene isoform mapping with the study of 3′UTR dynamics. Since there is no amplification step, SPIA significantly minimizes internal mis-priming that could generate false 3′ends during the cDNA preparation [132]. It is important to note that PAT-Seq relies on the binding affinity of *pab-1* to polyA tails of mature mRNAs, which are known to change in length in eukaryotic genes [161]. This in turn can create difficulties in the quantification of gene expression levels of libraries prepared with this technology. Although this is an inherent problem in all RNA-IP based approaches, our datasets correlated with previously published studies that did not use a PABPC-based approach (**Figure 2.12**) [119].

Using SPIA, we were able to build ~20,000 high-quality PAS clusters that allowed us to map 3′UTR ends at unprecedented resolution for ~6,000, ~2,000, and ~1,200 genes in intestine, pharynx, and body muscle datasets, respectively. Importantly, we found that more than 80% of these 3′UTR isoforms overlap with previously described datasets [88, 89], strongly suggesting that the vast majority of 3′UTRs detected by our approach are *bona fide* 3′ends of mRNAs.

~18-26% of genes use tissue-specific APA

We show that ~18-26% of total genes detected in intestine and muscle tissues used tissue-specific APA. Intestine genes seemed to favor *proximal-to-distal* PAS switches, leading to longer 3′UTR isoforms, while both muscle tissues alternated 3′UTR isoforms at a similar rate. Previous studies of 3′UTR datasets reported that as much as 46% of worm genes use APA across multiple tissues and developmental stages [88, 89]. This apparent discordance with our findings indicates that a large majority of genes in worms use only one 3′UTR isoform in a given tissue, suggesting that APA is indeed an important mechanism used by cells to regulate gene expression at the tissue-specific level.

Importantly, our work identified an overall high number of novel 3′UTR isoforms that were not present in past analyses [88, 89]. This pool spans from 9% to 16%, depending of the tissue examined (**Figure 2.16**). Previous work reported that the worm 3′UTRome is not saturated and other novel 3′UTR isoforms may be present [88]. Interestingly, the majority of the 3′UTR isoforms within this pool are tissue-specific (82% in intestine and 58% in the muscle), suggesting that perhaps these 3′UTR isoforms are also rare, and were not identified in earlier studies because of the limit of sensitivity of their mixed-tissue, transcriptome-wide approaches [88, 89].

Our analysis uncovered significant APA in worm tissues, but we could not identify upstream tissue-specific elements involved in 3′end formation, suggesting that in worms, other accessory tissue-specific factors [162] or their dosage [115] may play a role instead [88, 89].

Tissue-specific 3'UTR isoforms are linked to microRNA regulation

Past dogma that the protein and the transcription levels in cells are directly proportional is not accurate anymore [163, 164]. Thanks to the introduction of novel high-throughput technologies, it is now clear that there is not a direct correlation between the transcriptomes and the proteomes of cells or tissues. Instead, miRNAs, together with other ncRNAs and RNA binding proteins, play key roles in modulating the final gene output on its way to protein expression [96]. This modulation, when combined with the abundance of APA detected in this study, suggests a more complex picture, where there are not only negative regulatory networks through miRNAs, but also novel unexplored positive regulatory networks operated though APA. These positive networks are driven by genes that switch between 3′UTR isoforms to escape miRNA targeting, allowing their expression. In this view, both miRNAs and APA can, in principle, dramatically reshape gene expression output, implying they both play key roles in the establishment and maintenance of cell and tissue identity.

In this study we found that genes with tissue-specific 3′UTR isoforms are enriched in microRNA targets using both a 3-species (~49%) and a 5-species (~25%) conservation criteria. This was significantly more abundant than what we saw in randomly selected 3′UTR isoforms using a 3-species (25-40%) and a 5-species (10-19%) conservation criteria in each tissue. We also found a similar enrichment comparing our dataset to the experimentally validated ALG-1 footprints [127]. Our results in three worm somatic tissues link miRNA regulation to APA, showing that microRNA targets are much more abundant in ubiquitously expressed genes with tissue-specific 3′UTR isoforms than in genes that do not use APA.

Recently, microRNA populations from intestine and body muscle tissues were isolated using an RNA-IP strategy, providing a tissue-level atlas of microRNA expression [165]. Unfortunately, while this study suggests that miRNAs are also pervasive in worm tissues, it is still unclear which genes they target, and further experiments have to be performed to highlight these regulatory networks.

APAome.org: A resource for 3′UTR biology

We have compiled our tissue-specific transcriptomes into a useful online resource for scientists interested in 3′UTR biology and APA [128]. The APAome.org site uses an Apache web server and several custom-made Perl scripts that query a dedicated MySQL database. It is currently hosted in the Biodesign Institute at Arizona State University, and offers a simple and well-integrated interactive user interface to query gene records and 3′UTR isoform data, giving access to a dedicated gBrowse installation specifically designed to study APA in worms.

This database displays tracks for each tissue transcriptome, including tissue-specific APA, as well as curated 3′UTR data from previously published studies [88, 89, 119].

Experimental

Plasmids and molecular cloning

The PolyA-Pull plasmid was constructed adapting the Gateway pDONR221 (Invitrogen, Carlsbad, CA) as follows. The *pab-1* gene was amplified from N2 genomic DNA using a forward specific primer containing a SacII site, and a reverse specific primer containing BamHI and EcoRI sites (**Table 2.7**). The amplicon was then ligated *in-frame* with GFP (Marco Mangone, *unpublished*) using T4 DNA Ligase (NEB, Ipswich, MA) and SacII and EcoRI sites. The 3xFLAG epitope DNA sequence was obtained from the DNASU Plasmid Repository [166] (DNASU clone ID: HsCD00298297), and extracted using PCR amplification using a forward primer containing a BamHI site and reverse primer containing an EcoRI site. The amplicon was then ligated into pDONR221, (Invitrogen) downstream and *in-frame* with the *pab-1* gene using T4 DNA Ligase (NEB, Ipswich, MA). The Δ*pab-1*-Pull plasmid (GFP::Δ*pab-1*::3xFLAG), which does not contain the *pab-1* sequence and cannot bind polyA+ mRNAs, was prepared from the PolyA-Pull plasmid using the Stratagene QuikChange® Site-Directed Mutagenesis Kit following the manufacturer's guidelines (Stratagene, La Jolla, CA) (**Table 2.7**). The 3′UTR of the *unc-54* gene, cloned in Gateway pDONR P2R-P3 entry vectors [88], was used as an unspecific 3′UTR in all of the destination vectors in this study. The tissue-specific promoters were selected as the genomic sequence of DNA upstream of their transcription start site, up to 2kb. We have designed the primers using the UCSC Genome Browser and cloned the resultant amplicons from N2 genomic DNA into the Gateway™ pDONR P4-P1R entry plasmid (Invitrogen) (**Table 2.7**). We used Multisite recombination reactions (LR Clonase plus II, Invitrogen) to join the tissue specific promoters, the PolyA-Pull vector, and the *unc-54* 3′UTRs into the Gateway Compatible MosSCI destination plasmid pCFJ150 [131], (Addgene plasmid #19329), and used these vectors for the preparation of the transgenic strains.

Nematode strains and preparation of transgenic animals

Wild-type strain N2 worms were obtained from the CGC (University of Minnesota), which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). Worm strain EG4322 (to prepare MosSCI transgenics) were maintained at 16°C on HB101 containing NGM agar plates prior to microinjection [167]. Stable transgenic worm strains were prepared using the MosSCI technology as described [131].

Microinjection mixes consisting of pJL43.1(50ng/µl), pCFJ90(1ng/µl), pGH8(10ng/µl), pCFJ104(5ng/µl), and pCFJ150::TissuePromoter::GFP::*pab-1*::3xFlag::*unc-54* (25ng/µl) were microinjected into worm strain EG4322 (ttTi5605; *unc-119(ed9) III*), each of which was kindly provided by Priscilla Van Wynsberghe (Colgate University). Microinjection was carried out using a Leica DMI3000B microscope according to that described previously [131, 168]. Injected worms were plated on NGM growth media plates containing OP51 bacteria, and plates containing *unc-119* rescued (mobile) worms were chunked onto four new NGM plates and left to starve for at least 30 days at 25°C. Single dauer worms were plated onto small NGM plates, propagated for approximately 2 weeks, and verified for GFP expression using a Leica DMI3000B. DIC and fluorescent images were captured using a Leica DFC345FX mounted camera.

Worm gDNA extraction and MosSCI insertion verification

Genomic DNA was phenol-chloroform extracted from one full 60mm NGM plate from each transgenic worm strain, precipitated with sodium acetate, and washed in ethanol. To confirm the MosSCI integration of transgenes into the ttTi5605 intergenic region, we performed PCR using Standard Taq Polymerase (NEB, Ipswich, MA) using a forward primer annealing outside of the homologous flanking region (5′-CCTCTGAACTGGTACCTCA -3′) and a reverse primer annealing within the *unc-119* rescue cassette (5′- GGAAGAAGGAAAAGAGTGTGG -3′), both of which were provided by Priscilla Van Wynsberghe (Colgate University).

Western blotting

Western blotting for detection of the GFP::PAB-1::3xFLAG fusion protein in transgenic worms was carried out as follows. One full 60mm NGM plate of worms was washed with M9 media into a 1.5ml centrifuge tube and pelleted at 1500rpm. Worms were washed 2x in PBS Buffer and then resuspended in an equal volume of sample buffer (125mM Tris-Cl [pH 6.8], 4% SDS, 20% Glycerol, 0.5% bromophenol blue) supplemented with 10% beta-mercaptoethanol and boiled at 95°C for 5 minutes. The reaction was spun down and 15µl of supernatant was run at 200V on a 4-15% Tris-Glycine Criterion™ precast polyacrylamide gel (Bio-Rad, Hercules, CA) for 36 minutes. Electrophoretically separated proteins were transferred to an Amersham Hybond™-P blotting membrane (GE Healthcare, Little Chalfont, UK) using a Trans-Blot SemiDry Transfer Cell (Bio-Rad, Hercules, CA) at 23V for one hour. The membrane was blocked in blocking buffer (5% milk in PBS with 0.01% TWEEN-20) for 1 hour at room temperature followed by overnight incubation with ANTI-FLAG® antibody produced in rabbit (Sigma-Aldrich, St. Louis, MO). Following incubation, the membrane was washed 3x in blocking buffer and then incubated with a 1:1000 dilution of anti-Rabbit IgG HRP-linked secondary antibody (Cell Signaling, Danvers, MA #7074S) in blocking buffer for one hour. The membrane was finally washed 4x in PBST (1xPBS, 0.01% TWEEN-20) and then reacted with SuperSignal ELISA Femto Maximum HRP Substrate (Thermo Scientific, Rockford, IL), followed by imaging with a FluorChem FC2 Imager (Alpha Innotech, San Leandro, CA).

RNA immunoprecipitation

The mRNA tagging technique was adapted from past studies [127, 129]. Mixed-stage liquid worm cultures were grown as described [152] at 20°C. Approximately 10^6 *pab-1*::3xFLAG transgenic worms were harvested from liquid culture after 3 to 4 days, crosslinked for one hour in 0.5% paraformaldehyde in M9 solution, and flash frozen in ethanol-dry ice bath. Frozen pellets were crushed using a mortar and pestle in liquid nitrogen and the resulting frozen powder was transferred directly into lysis buffer (150mM NaCl, 25mM HEPES, pH 7.5, 0.2mM DTT, 10% glycerol, 0.0625% RNAsin, 1% Triton X-100), described in [169]. Total RNA was extracted from worm lysates using Trizol® Reagent (Life Technologies, Carlsbad, CA) and precipitated with isopropanol. An amount of lysate corresponding to 90μg of total RNA was added to 100μl of Anti-FLAG® M2 Magnetic Beads (Sigma-Aldrich, St. Louis, MO) and incubated overnight at 4°C. Each reaction was washed 3X in 200μl TBS and then 3X in 200μl Proteinase-K buffer with 1000 RPM mixing. Proteinase-K (4mg/ml) was added to the beads and incubated at 37°C for 30 minutes with 1000 RPM mixing. 7M Urea was added to beads and incubated at 37°C at 1000 RPM before RNA was extracted with Trizol® Reagent and precipitated with isopropanol and GlycoBlue (Ambion, Austin, TX). Precipitated RNA was treated with DNAse I (NEB, Ipswich, MA) for ten minutes and extracted again with Trizol® Reagent and isopropanol. RNA was resuspended in nuclease-free water and quantified using a Nanodrop® 2000c spectrophotometer (Thermo-Fisher Scientific, Waltham, MA).

RT-PCR and 3'RACE reactions

Precipitated RNA (50ng) was reverse transcribed with a $NVdT_{(23)}$ primer using SuperScript Reverse Transcriptase III (Thermo-Fisher Scientific, Waltham, MA). Three microliters of the reverse transcription reaction was used in each PCR reaction using Standard Taq Polymerase (NEB, Ipswich, MA) and primers specific to the 3′ end of each cDNA ORF (**Table 2.7**) or 3′UTR, as was the case for *unc-54* (forward primer sequence was extracted from previous publications [88]).

cDNA Library Preparation and sequencing

The eight cDNA libraries were prepared using at least 50ng of total RNA extracted from different tissues. We used the IntegenX's (Pleasanton, CA) automated Apollo 324 robotic preparation system to reverse transcribe RNA into cDNA and for DNA library preparation. The cDNA synthesis was performed using a SPIA (Single Primer Isothermal Amplification) kit (IntegenX and NuGEN, San Carlos, CA) [132]. Once the cDNA was generated, we assessed the quantity of the cDNA libraries using the Nanodrop instrument (Thermo). The cDNA Shearing was performed on a Covaris S220 system (Covaris, Woburn, MA). After the cDNA was sheared to approximately 300 base pair fragments, the Nanodrop instrument was used again to quantify the cDNAs in order to calculate the appropriate amount of cDNA necessary for library construction. Tissue-specific barcodes were then added to each cDNA library. The resultant 8 tissue-specific libraries were then pooled and sequenced using the HiSeq platform (Illumina, San Diego, CA) with a 2x100bp HiSeq run.

Bioinformatics analysis of RNA-Seq data

Raw Reads Mapping: Paired raw reads were demultiplexed by their unique tissue-specific barcodes and converted individually to FASTQ files by the CASAVA software (Illumina). Unique datasets were then mapped to the *C. elegans* gene model WS190 using the Burrows-Wheeler Aligner software (BWA) [170] with default parameters. A summary of the results produced by this approach is shown in **Table 2.1**. Mapped reads were further converted into a bam format and sorted using SAMtools software run with generic parameters [171].

Cufflinks/Cuffdiff Analysis: Expression levels of individual transcripts were estimated from the bam files by using Cufflinks software [172]. The fragment per kilobase per million base (FPKM) number was used to indicate the gene expression levels, and FPKM value >=1 was used as a threshold across all tissues profiled for defining expressed genes. The gene expression levels obtained in each tissue dataset were compared pairwise with other tissues using the Cuffdiff algorithm [172]. Cuffdiff algorithm detected 389 isoforms shared between pharynx and intestine, 286 between body muscle and intestine, and 175 between the two muscle tissues (p-value<0.05). Cufflinks was unable to assign an FPKM value for eight genes in our intestine dataset (*vit-5*, *rpl-24.1*, ZK484.1, *hmg-1.1*, *rps-12*, Y24D9A.8 , *rps-8* and *rpl-7A*). These genes were omitted in this study. The differential mRNA isoform analysis was performed with the CummeRbund package [173] using the output produced by the Cuffdiff algorithm. This analysis aimed to identify genes that change in expression level between tissues from large datasets. The data is displayed in **Figure 2.13** as a plot. We have detected between 175 and 389 tissue specific isoforms that have significantly different expression levels between two tissues. Tissue-specific unique genes were assigned if they have an FPKM>=1. Genes with an FPKM<1 were ignored in our analysis.

Comparison with other intestine datasets

A list of 3,502 genes present in the original *Haenni et al.* dataset was obtained from the supplementary materials section of the publisher, and used for our analyses. In addition, we downloaded from GEO and re-mapped the original raw 'sorted' BAM file used in the *Haenni et al.*, manuscript, using BWA [170] and Cufflinks [172] and standard parameters to the WS190 gene model. This mapping effort produced 5,840 clusters mapped to 3′UTRs of known genes with a FPKM >=1. This list was labeled "*Haenni et al., re-mapped*" and used for our analysis. The list of genes detected by *Pauli et al.*, and by *McGhee et al.*, was obtained from the supplementary materials accompanying their respective manuscripts [124, 136].

Gene expression localization and validation

We cloned the promoter region of eight randomly chosen genes, designing genome-specific primers that selectively amplify promoter regions spanning from -2,000nt to WS190-annotated start codon of the gene of interest. The results are shown in **Table 2.3**. The genes chosen for this analysis were C25A1.5, *nac-3p*, *tmd-2p*, *fat-2p*, *nas-1p*, *let-756p*, and *lin-3p.* These forward and reverse primers contain Gateway-compatible sequences to allow the cloning of the resulting promoter regions in Gateway-compatible entry vectors (**Table 2.7**). We used the GFP containing plasmid PolyA-Pull to drive GFP expression using these promoters. Each promoter was introduced at the 5′end of a MosSCI-compatible PolyA-Pull fused to the *unc-54* 3′UTR within the MosSCI-compatible destination vector pCFJ150 [174] using multisite Gateway recombination technology (Invitrogen). The finalized constructs were microinjected into young adult worms. At least two independent biological replicates per construct were screened for GFP expression. For each tissue, we defined a putative expression index, proportional to the FPKM values obtained for each gene in each tissue (* = FPKM < 100, **, FPKM = 100 – 200, *** FPKM > 200).

PolyA cluster preparation and polyA mapping

To map polyA-sites to WS190 worm annotations, raw sequence reads were filtered using custom made Perl scripts. We extracted reads containing greater than or equal to 30 consecutive adenine nucleotides at their 3′end. We obtained 14,472 total reads from intestine, 7,532 for pharynx, and 5,185 for body muscle (**Table 2.6**). The polyA elements were then removed and the reads were converted to FASTA format and aligned to the WS190 annotation using the Burrows-Wheeler Aligner [170] with standard parameters. Reads mapping to genomic regions containing >=65% adenosines in either direction and /or with less than 18 consecutively mapped nucleotides were discarded. The reads produced approximately ~27,000 high-quality PAS clusters mapped through the *C. elegans* genome. Each of these clusters was then bioinformatically attached to the closest gene within a 1,600nt range in the same orientation. To increase the stringency of our analysis, we ignored clusters with less than 5% of the total number of polyA reads detected for a given gene, and PAS clusters that mapped genomic regions with >40% adenosines, to eliminate as much background as possible. Each cluster had a median length of ~70nt with ~5x coverage, and mapped 3′UTRs of genes detected in the corresponding tissue with a FPKM>=1.

PAS analysis

Mapped polyA sites were compared with *Mangone et al.* and *Jan et al.* to map common 3′UTR isoforms between these datasets. We assigned common PolyA sites if the overlap was between +-10nt. PAS usage in **Figure 2.17** was calculated as in *Mangone et al.* [88]. PAS position and PAS nucleotide composition for 3′UTR isoforms in each dataset was extracted from *Mangone et al.* and used for the analysis in **Figure 2.17**.

PAS nucleotide frequency

We have bioinformatically extracted 70nt sequences between -50 and +20 from the cleavage site of all 3′UTR isoforms detected in each tissue, and used these sequences to plot the nucleotide frequency.

Promoter analysis

We have used custom Perl scripts to bioinformatically extract 600nt from -500 to +100 from genomic regions of genes in WS190, in our intestine, pharynx, and body muscle datasets. We then calculated and displayed the nucleotide frequency in the graph shown in **Figure 2.14**. This approach was used in the past by others to study promoter regions [119]. The analysis in **Figures 2.14 and 2.15** was performed binning these 600nt-long promoter regions in 100nt bins using custom Perl scripts (six bins total), and then calculating the frequency of all possible hexamer combinations in each bin. As a control, we have extracted genomic regions from a random set of genes. Each random dataset used in our analysis was composed of the same number of genes detected in each corresponding tissue. We then used custom Perl scripts to bin these regions in and search for enriched hexamers within each of these bins. ~70% of worm genes are trans-spliced at their 5′ends, make challenging to precisely identify worm promoter regions (Wormbook). The analysis in **Table 2.5** excluded promoter regions of genes present in operons.

Motif identification with MEME

Promoter regions from each tissue were subjected to analysis for enriched motifs using the MEME Suite [175]. We used the DREME tool to search for enriched short motifs (up to 8 bases) in the tissue-specific promoter datasets used in our promoter analysis, and performed a discriminative motif discovery search using different tissue datasets as negative controls. We then overlapped the motifs detected with DREME with the high quality transcription factor binding profile database JASPAR using the human and the worm datasets (version 2014) [176].

Transcription factor search analysis

We searched our tissue specific datasets for the presence of known transcription factors present in the wTF2.0 database [137], and compared the results with *Haerty et al.* [138].

Gene expression network visualization with Cytoscape

Tissue-specific genes, isoforms, and APAs were extracted from the data tables, reconfigured as a binary interaction format with three tissue types, and visualized as networks using Cytoscape v3.1 [177]. The FPKM values in the tissues were log2-transformed, converted to RGB color codes, and used to display relative expression levels among three tissues.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Table 2.7. Primers used in chapter 2.

CHAPTER 3

GENETIC VALIDATION OF THE CONSEQUENCES OF ALTERNATIVE POLYADENYLATION INVOLVING THE CAENORHABDITIS ELEGANS GENES TCT-1 AND RACK-1

Overview

Alternative polyadenylation (APA) is widespread in eukaryotic organisms, giving rise to a dynamic landscape of mRNA 3′ends [88, 89]. Despite nearly twenty years of research, it is still not understood how multiple 3′ends for the same gene are produced and the activities of the expressed 3′UTR isoforms are largely unknown. This process could, in theory, provide genes with a useful mechanism to modulate the activities of factors that regulate them given the multitude of regulators that target in 3′UTRs and dose gene expression.

Massive efforts to finely map the 3′ends of the *C. elegans* transcriptome revealed an especially broad landscape of 3′UTR isoforms that could dramatically impact the regulation of gene expression. While almost half of worm genes are subject to APA, their functions remain unclear. Given the pervasive role of gene regulation in metazoan tissue development and physiology, we and others have hypothesized that APA operates primarily at a tissue-specific level to drive gene regulatory mechanisms important for precisely coordinating those activities.

We have recently developed an approach, PAT-Seq (polyA-tagging and sequencing), that allows for isolation and sequencing of tissue-specific mRNA from small worm tissues [128]. This approach was applied to closely study the dynamics of APA in worm intestine, pharynx and body muscle tissues, revealing pervasive tissue-specific APA between these three somatic tissues. These studies also uncovered a particularly striking correlation between alternative polyadenylation and miRNA regulation that warrants further investigation. We uncovered a significant enrichment of bioinformatically and experimentally predicted miRNA targets in 3′UTRs of genes having tissue-specific 3′UTR isoforms. That is, the capacity for miRNA mediated gene regulation appears to be abundant in cases where unique APA events occur in each tissue. The data suggest that tissue-specific APA events could fine-tune gene regulation events through the coordinated addition and subtraction of miRNA targets between each tissue.

A particular limitation to this study is the general absence of precisely validated miRNA target information due to the general lack of experimental approaches. For this reason, we unfortunately do not have a resource that matches each candidate mRNA with a specific miRNA regulator. In absence of this information, bioinformatically predicted miRNA targets are commonly used. These software use algorithms that incorporate metrics known to influence miRNA targeting mechanistics, such as nucleotide conservation and the degree of base-pairing between the miRNA and its target [57, 178]. However, they are unfortunately hampered by the tendency to over-predict, leading to large numbers of false-positive and negative target predictions [53]. Thus, bioinformatics miRNA target prediction software are useful for initial inquiries, but experimental validation remains the gold standard for miRNA genetics research.

Our results in *C. elegans* are not unique. Several groups have recently uncovered correlations between APA and miRNA regulation in normal human tissue and in malignant cell lines [96, 111]. However, it is difficult to identify the consequences of APA on miRNA induced gene regulation in these systems due to the lack of context provided by independent cell lines grown outside of the organism. Using *C. elegans* as a model system, we are uniquely positioned to validate interactions between tissue-specific APA events and post-transcriptional gene regulation induced by miRNAs within the living organism. The array of genetics approaches coupled with their rapid development will enable unique insights into not only APA and miRNA regulation, but also its biological consequences.

Together, our data suggest that APA may influence miRNA induced post-transcriptional gene regulation by controlling interplay between the gene and its miRNA target uniquely in each somatic tissue. Why might cells need to allow genes to escape regulation by miRNAs as opposed to directly down-regulating expression of the miRNA? Hypothetically, the answer to this question lies in the dynamics of miRNA mediated gene regulation known to occur. miRNAs commonly regulate more than one gene [179], [180] and often operate in networks where the same gene harbors targets for multiple miRNAs. While the miRNAs are commonly expressed across multiple tissues, many of the genes targeted in the network may require regulation in one tissue, but not in the other. APA could provide a mechanism to allow the same gene to uniquely escape regulation by the network in the tissue where it needs to be expressed.

Our application of PAT-Seq in *C. elegans* intestine and muscles has uncovered many examples of genes that lose predicted miRNA targets due to APA induced 3′UTR shortening in one tissue but not in another. The data suggest these genes may do so to escape miRNA mediated gene regulation and, in doing so, dose their expression to levels required to maintain a phenotype associated with that tissue's function. An intriguing example we chose to investigate in detail involves the ubiquitously expressed worm genes *rack-1* and *tct-1*.

rack-1 is the *C. elegans* ortholog of human Rack-1 (receptor for activated C-kinase). This widely conserved eukaryotic gene encodes a seven-bladed multifunctional scaffolding protein shown to be important for a variety of biological processes in many different cell types [181].Its name originates from its early identification as a major effector in an Akt-signaling cascade that recruits the C-kinase to active ribosomes on targeted mRNAs, enabling C-kinase mediated phosphorylation of translation machinery and stimulation of translation elongation [182].

Rack-1 is expressed in many different cell types [181] [183] and is involved with an array of biological processes in each, suggesting its modulation may be a key factor in supporting their specific activities. For example, in smooth muscle cells, Rack-1 signals calcium release from the endoplasmic reticulum required for muscle cell proliferation [184]. Additionally, Rack-1 is commonly modified at the post-translational level to control its activities [11], [185]. The induction and expression levels of Rack-1 are tightly regulated and misregulation of Rack-1 expression levels are associated with cancers, brain disorders and muscle atrophy [181]. While *C. elegans rack-1* has been less studied in general, it was recently implicated as a critical factor in the negative regulation of miRNA biogenesis [186]. A genetics approach showed that *rack-1* depletion increases levels of the *let-7* miRNA and disrupts terminal cell differentiation normally controlled by this miRNA. Together, these data suggest that the precise regulation of *C. elegans rack-1* may in fact be very important for supporting its functions in specific tissues and highlighting the potential need for this gene to use APA to modulate its expression on a tissue-specific basis.

C. elegans tct-1 is the ortholog of human TPT-1 (translationally controlled tumor protein). In humans, TPT-1 was originally characterized as a downstream target of the p53 transcription factor [187] and later identified as a potent regulator of cell fate upon the observation that reduction in its expression levels could revert leukemia cells to their original non-malignant state [188].

Interestingly, translation of this gene is directly controlled by the nutrient sensing mTOR pathway [189], which is also involved in controlling ribosome activities [190]. Thus, *tct-1* may be functionally related to *rack-1*. TPT-1 is pleiotropic and is suggested to have a powerful role in cell fate decisions by guiding proteins that control cell proliferation, inflammation, DNA damage response, and ribosome biogenesis [187]. Further, TPT-1 positively regulates RhoA activities in smooth muscle cells. Therefore, regulation of TPT-1 expression may have a powerful impact on cell fate decisions in muscle cells [191].

Less is known about the function of *tct-1* in nematodes. So far, a single study has investigated its role using a reverse genetics approach finding that *tct-1* is required for proper egg laying [192], which is in part mediated by the function of *C. elegans* body muscle. Given the importance of egg laying to worm generational and reproductive viability, these data suggest that the precise regulation of *tct-1* in worms may be important.

Here, we investigate the dynamics of tissue-specific APA events between body muscle and intestine tissues involving the genes *rack-1* and *tct-1* in *C. elegans*. We biochemically validated the short 3′UTR isoform of both genes as expressed in body muscle, while an alternate cleavage event expresses the long form in the intestine. Using a unique vector-based tool having dual fluorochrome reporters, we forced expression of the long 3′UTR isoform of both genes in the body muscle tissue of transgenic worms and observed repression for each case. Removing the miRNA targets in the distal portion of the forced long 3′UTR rescued sensor expression for both cases. We applied reverse genetics approaches (RNAi) to study the role of *rack-1* and *tct-1* in the body muscle, finding their expression is required for normal locomotion. Together, these data suggest a critical role for APA in allowing *rack-1* and *tct-1* to escape the *miR-50* mediated post-transcriptional gene regulatory network, enabling their local cellular functions in body muscle required for locomotion.

Results

A body muscle-specific APA event enables expression of rack-1 and tct-1 with short 3′UTR isoforms

Our application of PAT-Seq to study the dynamics of APA in *C. elegans* intestine, pharynx and body muscle tissues revealed many examples of differential 3′end expression for the same genes. We carefully studied patterns of APA between each tissue and identified two genes, *rack-1* and *tct-1*, that display an interesting dynamic (**Figures 3.1 and 3.2**). The worm ortholog of the receptor for activated C-kinase, *rack-1*, was mapped with a short 3′UTR isoform spanning 41 nucleotides from the STOP codon to the poly(A)-tail in the body muscle (**Figure 3.1**).

We mapped the alternate APA event in the pharynx muscle and intestine tissues, where the distal PAS element was used, leading to expression of the long 3′UTR isoform spanning 78 nucleotides from the STOP codon to the poly(A)-tail for this gene (**Figure 3.1**)

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.1. Mapped poly(A)-sites for *rack-1* in intestine and body muscle tissues. Screenshot image of the 3′end of the *rack-1* gene from the gBrowse interface of APAome.org. We mapped a single polyA-cluster in the body muscle (blue peak) corresponding to a short 3′UTR isoform and two polyA-clusters in the intestine (grey) corresponding to both short and long 3′UTR isoforms expressed in this tissue.

Similarly, we mapped a short 3′UTR isoform for the worm ortholog of translationally controlled tumor protein, encoded by the gene *tct-1*, expressed in the body muscle (**Figure 3.2**). The short isoform spans 51 nucleotides from the STOP codon to the poly(A)-tail. We mapped a longer 3′UTR isoform in the pharynx muscle and intestine that spans 92 nucleotides from the STOP codon to the poly(A)-tail (**Figure 3.2**). Interestingly, we also mapped a small poly(A)-cluster corresponding to the longest 3′UTR mapped for this gene in the worm 3′UTRome that spans 113 nucleotides (**Figure**

3.2). Together, these sequencing data suggest that *rack-1* and *tct-1* are expressed with short 3′UTR isoforms in the body muscle due to a unique APA event in this tissue.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.2. Mapped poly(A)-sites for *tct-1* in intestine and body muscle tissues. Screenshot image of the 3′end of the *tct-1* gene from the gBrowse interface of APAome.org. We mapped a single polyA-cluster in the body muscle (blue peak) corresponding to a short 3′UTR isoform and two polyA-clusters in the intestine (grey) corresponding to both short and long 3′UTR isoforms expressed in this tissue.

We focused on the APA dynamics of *rack-1* and *tct-1* in the body muscle, where the short 3′UTR isoforms are expressed, and in the intestine where abundant expression of the long 3′UTR isoforms for both genes have been mapped. Although we used several approaches to build high-quality poly(A)-clusters in these tissues used to define the 3′ends of each gene, we sought to validate the expression of the mapped 3′ends using an alternate approach. Using 3′Rapid Amplification of cDNA ends (3′RACE), we targeted the 3′ends of *rack-1* and *tct-1* mRNA and amplified their 3′UTRs (see Experimental, **Figures 3.3 and 3.4**). The results for *rack-1* indicate abundance of the short 3′UTR isoform in both tissues (*blue arrow*, **Figure 3.3**). However, the long 3′UTR isoform was only detected in the intestine, suggesting that cleavage at the distal PAS element was specific to this tissue (*red asterisk*, **Figure 3.3**). We observed long and short 3′UTR isoforms of *tct-1* expressed in intestine and body muscle tissues (**Figure 3.4**). However, in the intestine we observed preferential expression of the long 3′UTR isoform and conversely, abundant expression of the short 3′UTR isoform for this gene in the body muscle (**Figure 3.4**).

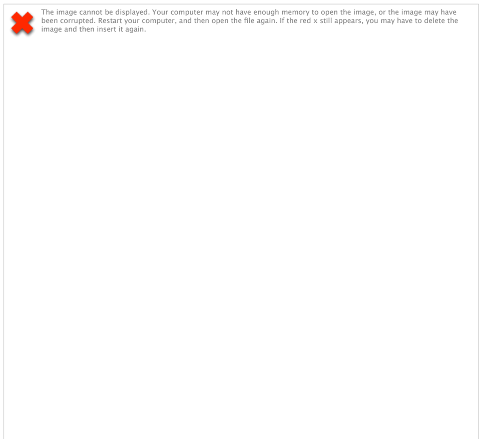

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.3. Biochemical validation of *rack-1* tissue-specific 3′UTR isoform expression. We used a 3′RACE strategy to amplify the 3′end of *rack-1* from intestine (int) and body muscle (bm) mRNA preparations. The short 3′UTR isoform (blue arrow) was detected in both tissues, while the long 3′UTR isoform (red asterisk) was uniquely detected in the intestine.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.4. Biochemical validation of *tct-1* tissue-specific 3′UTR isoform expression. We used a 3′RACE strategy to amplify the 3′end of *tct-1* from intestine (int) and body muscle (bm) mRNA preparations. The short 3′UTR isoform (blue arrow) was more abundant in the body muscle, while the long 3′UTR isoform (red asterisk) was abundant in the intestine tissue.

Together these data cross-validate our results from poly(A)-cluster mapping of *rack-1* and *tct-1* and support a model where each gene is expressed with short 3′UTR isoforms in the body muscle and long 3′UTR isoforms in the intestine.

rack-1 and tct-1 use APA to escape miRNA regulation in the body muscle

The human homolog of worm *rack-1* is expressed in many different cell lines and their expression levels are precisely dosed to support a range of different activities between cell types [181] [183]. Reflecting these data, *rack-1* is also expressed in most worm tissues [193]. Our data so far indicate that *tct-1* is also abundantly expressed in the pharynx, intestine, and body muscle [128]. We therefore hypothesized that the tissue-specific APA events we observed between intestine and body muscle tissues could modulate *rack-1* and *tct-1* expression levels between each tissue, perhaps by controlling their interactions with miRNAs.

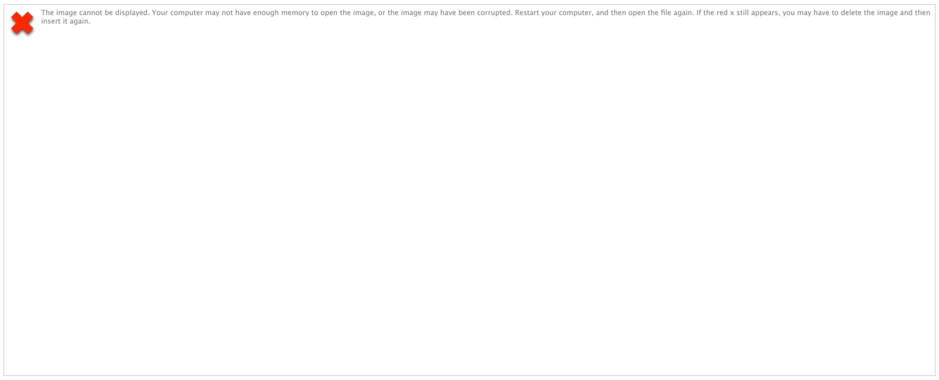
We used PicTar miRNA target prediction software to search for potential targets in *rack-1* and *tct-1* 3′UTRs. This analysis revealed predicted targets for *miR-50*, a ubiquitously expressed miRNA, in both genes (**Figure 3.5**). *miR-50* is predicted to target the most distal portion of the long 3′UTR isoform of *rack-1*, nearest the distal PAS element that is expressed in intestine (**Figure 3.5**). The same miRNA is predicted to target both of the longest 3′UTR isoforms of *tct-1* that are also expressed in the intestine (**Figure 3.5**). Notably, PicTar software also predicts a miRNA targets for *miR-85* in the longest 3′UTR isoform of *tct-1* and *rack-1* (**Figure 3.5**). Importantly, no miRNA targets are predicted in the sequence region expressed in the short 3′UTR isoform of both genes that we have validated as expressed in the body muscle tissue (**Figure 3.5**). PicTar also predicts a *miR-50* target in another gene, *pek-1*, that is only expressed with one 3′UTR isoform suggesting it is part of the same network targeted by *miR-50* (*data not shown*).

Together, these data suggest that *tct-1* and *rack-1* may escape *miR-50* and *miR-85* in the body muscle through APA-induced 3′UTR shortening.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.5. *miR-50* and *miR-85* network involving *rack-1* and *tct-1*. Cartoon illustrating all 3′UTR isoforms mapped for *rack-1* (blue) and *tct-1* (orange) in the worm 3′UTRome. We mapped the short and long 3′UTR isoforms of these genes in the body muscle (bm) and the intestine (int), respectively. The relative location of PicTar predicted *miR-50* and *miR-85* targets in the distal region of the long 3′UTR isoforms are indicated.

To investigate our hypothesis that *rack-1* and *tct-1* escape miRNA regulation in the body muscle, we developed a vector-based sensor tool, called pAPAreg, that can be used to sense post-transcriptional gene regulation by miRNAs in transgenic worms (**Figure 3.6**). pAPAreg contains two fluorochromes, mCherry and Green Fluorescent Protein (GFP), separated by a *trans*-spliceable element (**Figure 3.6**). This *trans*-splicable element is derived from the well-characterized sequence region between *trans*-spliced genes *gpd-1*/*gpd-2* in the *mai-1* operon [194]. We placed the body muscle-specific promoter from the gene *myo-3* upstream of this construct, allowing transcription of a polycistronic pre-mRNA followed by SL2 *trans*-splicing between mCherry and GFP. The mCherry fluorochrome therefore serves as a transcriptional reporter since it is spliced away from GFP and it is not subject to conditions between experiments. However, the GFP is expressed with a 'test' 3′UTR containing putative miRNA targets and therefore reports the translation level (**Figure 3.6**). Importantly, this tool includes a degron tag placed downstream of GFP to limit its protein stability and allow detection of sometimes subtle repressive events mediated by miRNAs.



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.6. The pAPAreg expression construct used to detect miRNA-mediated gene repression in 3′UTRs. A tissue-specific promoter drives expression of the mCherry-GFP-PEST operon cassette. After *trans*-splicing, mCherry is translated, while GFP-PEST, is subject to regulation dictated by miRNA targets (purple asterisk) in the 3′UTR placed downstream of it. Deletion of PAS1 allows expression of the long 3′UTR isoform.

We have used this vector tool to 'force' expression of the long 3′UTR isoforms of *rack-1* and *tct-1* in the body muscle and sense putative post-transcriptional gene regulation induced by the *miR-50* or *miR-85* miRNAs (**Figures 3.7 and 3.8**). Expression of wild type *rack-1* 3′UTR resulted in unabated expression of GFP throughout worm developmental stages (**Figure 3.7 i**). We forced expression of the long 3′UTR isoform of *rack-1* by deleting the proximal PAS element used to express the short 3′UTR in body muscle and observed global repression of GFP (**Figure 3.7 ii**). GFP expression was significantly rescued when we expressed the long 3′UTR isoform containing a deletion of the predicted *miR-50* target (**Figure 3.7 iii**). We also observed a significant rescue in GFP after deleting the predicted *miR-85* target from the distal portion of the long 3′UTR isoform (**Figure 3.7 iv**). Taken together, these results provide evidence that *rack-1* escapes *miR-50* and *miR-85* induced repression from a body muscle-specific APA event.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.7. *rack-1* escapes *miR-50* and *miR-85* regulation in the body muscle through APA. *Top:* Representative mCherry and GFP fluorescent images of transgenic worms expressing pAPAreg in the body muscle using the *myo-3* promoter and the *rack-1* 3′UTR with (i) wild type sequence (wt), (ii) deleted PAS1 (ΔPAS1), (iii) deleted PAS1 and *miR-50* target (ΔPAS1;Δ*miR-50*), or (iv) deleted PAS1 and *miR-85* target (ΔPAS1;Δ*miR-85*). *Bottom*: Quantification of GFP fluorescence intensity relative to mCherry for each *rack-1* transgenic worm line pictured in left panel using Image-J software (n=34, ****=p<0.0001, paired T-test).

We then tested the effect of forcing the long3′UTR isoform of *tct-1* in the body muscle (**Figure 3.8**). Similar to our results expressing the wild type *rack-1* 3′UTR, wild type expression of *tct-1* allowed GFP expression through all developmental stages with no obvious changes in expression levels (**Figure 3.8 i**). We observed a significant decrease in GFP expression when forcing the expression of the long 3′UTR isoform of *tct-1* after deleting its proximal PAS element (**Figure 3.8 ii**).
Expression levels of GFP were significantly rescued when expressing the long 3′UTR isoform of *tct-1* after deleting the predicted *miR-50* target in the distal portion of the 3′UTR (**Figure 3.8 iii**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.8. *tct-1* escapes *miR-50* regulation in the body muscle through APA. Left: Representative mCherry and GFP fluorescent images of transgenic worms expressing pAPAreg in the body muscle using the *myo-3* promoter and the *tct-1* 3′UTR with (i) wild type sequence (wt), (ii) deleted PAS1 (ΔPAS1), or (iii) deleted PAS1 and *miR-50* target (ΔPAS1;Δ*miR-50*). Right: Quantification of GFP fluorescence intensity relative to mCherry for each *tct-1* transgenic worm line pictured in left panel using Image-J software (n=27, ****=p<0.0001, paired T-test).

In summary, we developed a useful vector based tool that enables ‘sensing’ of post-transcriptional gene regulation in transgenic worms (**Figure 3.7**) and have used this vector tool to provide evidence that the worm genes *rack-1* and *tct-1* use body muscle-specific APA events escape miRNA regulation by the ubiquitously expressed *miR-50* (**Figure 3.8**).

rack-1 and tct-1 are important for overall viability and locomotion

Our results are consistent with the emerging hypothesis in the field that APA drives combinatorial variation between *cis*-elements in the 3′UTR and the *trans*-acting miRNAs that target them allowing tissue-specific gene dosing events where miRNAs are otherwise co-expressed (**Figure 3.9**- hypothesis). We hypothesized that *rack-1* and *tct-1* may have body muscle-specific functions that require them to escape post-transcriptional regulation by miRNAs in this tissue to sufficiently dose their expression.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.9. Hypothesis: APA rearranges the *miR-50* targeting network in the body muscle to support its tissue-specific functions.

Based on this hypothesis, we reasoned that *rack-1* and *tct-1* may be especially important for specialized functions in the body muscle, such as locomotion. We used an RNAi approach to identify the importance of both genes. Knockdown of *rack-1* resulted in ~40% embryonic lethality (**Figure 3.10**) indicating its importance for overall viability and early development. Of worms that bypass embryonic lethality, ~30% are defective in locomotion or display uncoordinated phenotypes suggesting *rack-1* is overall important for locomotion (**Figure 3.11**).

Similarly, knockdown of *tct-1* resulted in embryonic lethality at a rate of ~10% (**Figure 3.10**) and ~8% of larval worms display an uncoordinated phenotype (**Figure 3.11**). Together, these results suggest that *rack-1* and *tct-1* may be important for supporting body muscle functions in *C. elegans*, giving reason to use of APA to counteract miRNAs in this tissue and allow their expression.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.10. RNAi mediated knockdown of *rack-1* and *tct-1* results in partial embryonic lethality. RNAi by feeding experiments performed in L4 animals confer embryonic lethality to F0 generation worms at ~40% and ~10% penetrance for *rack-1* and *tct-1* respectively, (n=10). Feeding with *par-2* (RNAi) was used as a control. (-), treated with null(RNAi).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 3.11. Worms fed *rack-1* or *tct-1* RNAi exhibit uncoordinated locomotion. Shown are the results from larval worms that bypassed embryonic lethality, shown above (n=10). *rack-1*(RNAi) resulted in ~30% uncoordinated phenotypes and *tct-1*(RNAi) resulted in ~8% uncoordinated phenotypes. (-), treated with null(RNAi).

Discussion

Evidence for multiple 3'UTR isoform expression of rack-1 and tct-1 within tissues

We have used a biochemical strategy, 3′RACE, to validate the expression of tissue-specific 3′UTR isoforms of *rack-1* and *tct-1* from mRNA immunoprecipitated from *C. elegans* intestine and body muscle. This approach has shown that 3′UTR isoforms are largely not expressed in an absolute manner. Rather, we detected changes in the relative abundances of multiple 3′UTR isoforms expressed for both genes. Other groups have also reported that APA most often changes the frequency of PAS element usage, thereby modulating the ratio of 3′UTR isoforms expressed for the same gene. Our 3′RACE results for *rack-1* suggest that the short 3′UTR isoform of this gene is actually the most abundantly expressed in the intestine. However, the long 3′UTR isoform is not detected at all in the body muscle suggesting that the distal PAS is only used in the intestine.

This difference in ratios could reflect dynamics of APA within specific tissues during development. For example, early larval stage worms may preferentially express the short 3′UTR of *rack-1* in the intestine to promote its expression required for activities supporting tissue development, while the expression ratio shifts in preference of the long 3′UTR in adults repress its activities when no longer needed. Further experiments, such as staged tissue-specific mRNA pull-downs, will have to be performed to address this question.

Alternatively, the ratio of 3′UTR expression may control gene expression in the same special and temporal context by limiting the effect of miRNA targeting on the total transcriptional product. For example, low levels of expressed long 3′UTR isoforms that bear miRNA targets coupled with highly expressed short 3′UTR isoforms for the same gene may allow a precise fine-tuning of gene expression. Evidence from biochemical studies of 3′end formation dynamics have suggested that gene expression can be controlled more directly by modulating the efficiency of cleavage and polyadenylation [75].

APA as a mechanism to escape miRNA regulation in C. elegans

Here, we have provided evidence that the *C. elegans* genes *rack-1* and *tct-1* are subject to a specific APA event in the body muscle that induces expression of short 3′UTR isoforms and allows them to escape repression by miRNAs. So far, this idea has only been genetically validated for a single gene. In murine quiescent satellite cells (QSCs), the differentiation-inducing transcription factor Pax3 is maintained at low levels due to the broad expression of miR-206, a miRNA that targets the distal portion of its 3′UTR isoform[97]. In a subset of QSCs, APA of Pax3 enables miR-206 target exclusion, allowing it to counteract miRNA repression and dose its expression to levels sufficient to induce muscle cell differentiation[97]. This example highlights an important positive regulatory role for APA in the post-transcriptional control of gene expression. However, it is not clear if this is an isolated example or if APA is instead a conserved mechanism allowing genes to escape miRNA regulation.

Our results for *rack-1* and *tct-1* suggest APA also allows genes to escape regulation by miRNAs in *C. elegans*. These results, together with the widespread and tissue-specific expression of alternative polyadenylation in *C. elegans* and the enrichment of miRNA targets in affected genes suggests an extensive positive regulatory role for APA in worms where genes could rearrange miRNA targeting networks to precisely dose their expression between tissues.

Given the ubiquitous expression of *rack-1* and *tct-1* among worm tissues and their established roles in mammalian cells [181] [183], both genes likely play an important role in the body muscle tissue of *C. elegans*. We have used a reverse genetics approach to provide evidence that both genes appear to be important for locomotion, indicating their activities in body muscle. However, this approach is not ideal since RNA from both genes is depleted transcriptome-wide and our phenotypes cannot be distinguished from loss of function in other tissues that support locomotion, such as neurons. Indeed, a role for *rack-1* in promoting axon guidance during neural development has already been shown [194]. Further experiments, such as body muscle-specific rescue of *rack-1* expression in mutants carrying a defective *rack-1* allele, are necessary to provide evidence for a cell autonomous role of each gene in supporting body muscle activities.

Experimental

3'RACE Reactions

For each N2 total RNA or intestine or body muscle-specific RNA sample (from RNA-IP of *ges-1*::PAP and *myo-3*::PAP worms, [128]), 100ng RNA was reverse transcribed with an NVdT$_{(23)}$ primer containing a 5′anchor sequence [195] (**Table 3.1**) using Superscript

Reverse Transcriptase III (Thermo-Fisher Scientific), according to the manufacturer's protocol. One microliter of the resulting cDNA was used as a template for each PCR reaction, along with 1μM of gene-specific forward primer for *rack-1* or *tct-1* (**Table 3.1**), the anchor reverse primer (**Table 3.1**) and Taq DNA Polymerase (NEB) to drive the reactions.

Nematode strains and transgenesis

The EG4322 background worm strain, which we used to prepare pAPAreg expressing transgenic worm lines, were maintained at 16°C on NGM plates seeded with HB101 bacteria prior to microinjection. Extrachromosomal array transmitting worm lines were prepared by microinjecting pCFJ150 Tissue-specific Promoter::pAPAreg::3′UTR (25ng/μl) along with the unspecific plasmid pMA122 into the background worm strain EG4322 (ttTi5605; *unc-119*(*ed9*) *III*), which were kindly provided by Priscilla Van Wynsberghe (Colgate University, Hamilton, NY, USA). Microinjection was carried out as described previously [168] using a Leica DMI3000B microscope.

Preparation of 3′UTR entry plasmids

We cloned the *rack-1* and *tct-1* 3′UTR sequences from the annotated STOP codon of each gene to 35 nucleotides downstream of the most distal annotated poly(A) signal element (for *tct-1*) and 73 nucleotides downstream of the most distal annotated poly(A) signal element (for *rack-1*).

Each sequence was amplified from N2 worm genomic DNA using forward primers overlapping the STOP codon and a reverse primer targeting the region downstream of the poly(A) signal element designed using the UCSC genome browser (**Table 3.1**). We used BP clonase™ reactions (Invitrogen) to clone each 3′UTR sequence into the third position Gateway™ pDONR P2R-P3 entry plasmids.

To prepare 3′UTR plasmids that would 'force' expression of the long 3′UTR isoforms for each gene we deleted their mapped proximal PAS elements from each pDONR P2R-P3 DONOR plasmid and replaced this element with the BglII restriction site to maintain the 3′UTR size. We replaced the *rack-1* 3′UTR proximal PAS 'AATAAA' located 28 nucleotides from the STOP codon in the pDONR P2R-P3 *rack-1* plasmid using the Stratagene QuikChange® Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA, USA) and the *rack-1delPAS* primers (**Table 3.1**). The proximal PAS 'AATGAA' in the *tct-1* 3′UTR was replaced with a BglII restriction site using the Stratagene QuikChange® Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA, USA) and the *tct-1delPAS* primers (**Table 3.1**).

We deleted select predicted miRNA targets from the distal portion of each 3′UTR in their pDONR P2R-P3 DONOR plasmids using the Stratagene QuikChange® Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA, USA) and the forward and reverse primers specified in **Table 3.1**.

Plasmids and molecular cloning

Molecular cloning of the Gateway™ pDONR P4-P1R entry plasmids containing the body muscle-specific *myo-3* promoter have been described previously [128], and were used in this research with no modifications.

The pDONR ROG plasmid was prepared joining the mCherry sequence, a *trans*-splicable region between *gpd-2* and *gpd-3* and the GFP sequence in the pDONR221 vector backbone. The restriction sites used were introduced into pDonr221 using the Stratagene QuikChange® Site-Directed Mutagenesis Kit following the manufacturer' s guidelines (Stratagene, La Jolla, CA, USA). All primers used in this study are shown in **Table 3.1**. We used two different versions of the pAPA_reg cassette in this study, named APAreg_1 and APAreg_2.

To prepare the pDONR 221 APAreg_1 entry plasmid, we amplified the PEST sequence from pAF207 [196], kindly gifted by Allison Frand, using a forward primer containing AgeI restriction sites and a reverse primer containing KpnI sites (**Table 3.1**). We added AgeI and KpnI restriction sites downstream of GFP in the pDONR ROG plasmid using the nROGinsAgeIKpnI primers (**Table 3.1**) and used them to ligate the amplified PEST sequence downstream and *in frame* of GFP in the pDONR 221 ROG entry plasmid using NEB Quick ligase (NEB, Ipswich, MA). We observed slightly stronger GFP expression with the pDONR 221 APAreg_1 vector and used it in the experiments with the *tct-1* 3′UTR.

The pDONR 221 APAreg_2 entry plasmid contains the *rpl-10* CDS sequence upstream and in frame of the mCherry and the GFP ORFs, to increase the vector's nuclear localization. We first added an EcoRI restriction site to pDONR 221 ROG upstream of the mCherry using the Stratagene QuikChange® Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA, USA) using the mCherry_ins_EcoRI primer and a ClaI restriction site downstream of GFP using the GFP_insClaI primer (**Table 3.1**). The *rpl-10* sequence was amplified from N2 worm genomic DNA using a forward primer containing a SpeI restriction site and a reverse primer containing an EcoRI restriction site. The amplicon was then ligated upstream and in-frame of the mCherry sequence in pDONR ROG using NEB Quick ligase (NEB, Ipswich, MA). To ligate *rpl-10* upstream of and in-frame with GFP, the *rpl-10* sequence was amplified from N2 worm genomic DNA using a forward primer containing a SacII restriction site and a reverse primer containing a ClaI restriction site (**Table 3.1**). The amplicons were then ligated into pDONR 221 ROG upstream of and *in-frame* with GFP coding sequences using NEB Quick ligase (NEB, Ipswich, MA).

To prepare the pAPAreg expression vectors, we joined the *myo-3* promoter, each Gateway™ cassette and the 3′UTR of interest (see preparation of 3′UTR entry plasmids above) into the Gateway™ compatible destination plasmid pCFJ150, which contains the *unc-119* rescue cassette, using Multisite recombination reactions (LR clonase plus II, Invitrogen).

Nematode imaging and fluorescence quantification

The fluorescence produced by extrachromosomal array worms carrying the pAPAreg expression plasmid transgene was detected using a Leica DMI3000B microscope. Images were captured using a Leica DFC345FX mounted camera with Gain=1X, Gamma=0.5 and 1 second exposure. GFP/mCherry fluorescence ratios were quantified using the integrated density (ID) function of ImageJ software [197] using the formula $[(ID_{GFPt}-((ID_{GFPb}/Area_b)xArea_t)]-ID_{GFP\_N2}] / [(ID_{mCherry\_t}-((ID_{mCherry\_b}/Area_b)xArea_t)-ID_{mCherry\_N2}]$ where: ID_{GFPt} and $ID_{mCherry\_t}$ are the integrated density values of each transgenic worm image obtained from GFP and mCherry channels, respectively. ID_{GFPb} and $ID_{mCherry\_b}$ are the integrated density values obtained from a small selection of the dark area (background) surrounding the worm in each image. $Area_b$ is the area of this small background selection and $Area_t$ is the total area of the entire image. ID_{GFP_N2} and $ID_{mCherry\_N2}$ are the average integrated density values obtained from non-fluorescent N2 worms (n=15) in the GFP or mCherry channels, respectively.

RNA interference assays

Each RNAi clone was obtained from the Julie Ahringer library [198] and the RNAi by feeding procedure was performed as described [199]. Briefly, each RNAi clone was grown in LB overnight at 37°C, 1,000rpm. Each clone was plated on small NGM media plates supplemented with 1mM IPTG and activated overnight at room temperature. To observe embryonic lethality and uncoordinated phenotypes, ten L4 stage *myo-3*::PolyApull expressing transgenic worms [128] were plated onto the seeded plates and incubated at 20°C for 24 hours. We then singled the adult worms onto new plates seeded with the same RNAi clones. After incubation at 20°C for 12 hours, the adults were removed and larval offspring on the plates were allowed to incubate for 24 hours before scoring their phenotypes.

Table 3.1. Primers used in chapter 3.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

CHAPTER 4

A ROLE FOR ALTERNATIVE POLYADENYLATION IN REARRANGING TISSUE-SPECIFIC MIRNA REGULATORY NETWORKS IN GENES COMMONLY EXPRESSED AMONG EIGHT CAENORHABDITIS ELEGANS SOMATIC TISSUES

Overview

Multicellular organisms rely on sophisticated gene expression programs to confer tissue identity and maintain homeostasis. The mRNA molecule is a dynamic mediator of these programs as it is capable of transferring genetic information into many different isoforms that shape gene expression outputs and precisely direct protein expression. However, the dynamics of mRNA expression in the somatic tissues of living organisms that give rise to their specialized functions are still not clear in most cases. Thus, mapping metazoan transcriptomes at the tissue-specific level to identify gene isoforms and their expression levels is key to understanding the how mRNA coordinates development in normal states and how its expression is disrupted in disease.

The small nematode *C. elegans* is useful for such studies, since it has a complete cell-lineage map [102], its development is well studied at the physiological and molecular level [6, 7], it has small, relatively simple tissues, and its transcriptome has been extensively characterized [101] [104]. The development of biochemical approaches to isolate tissue-specific mRNA have been applied to study a range of tissue transcriptomes in worms, from the large intestine [119] (McGhee et al. 2007) [124], to smaller tissues composed of just a few cells, such as sensory neurons [126]. However, most of these approaches have limited resolution as they have classically relied on low-throughput technologies, such as microarrays or tiling arrays, for detection.

We have recently developed a method, called PAT-Seq, which is an adaptation of the mRNA-tagging method coupled with high-throughput sequencing that improves the resolution of tissue-specific transcriptome profiling in worms [128]. In this method, transgenic worms expressing a 3xFLAG-tagged cytoplasmic poly(A)-binding protein (PABPC) using tissue-specific promoters in each tissue of interest are prepared, followed by crosslinking and immunoprecipitation of tissue-specific mRNA. Our application of PAT-Seq to the *C. elegans* intestine, pharynx, and body muscles allowed the mapping and study of thousands of expressed mRNAs and mRNA isoforms that are dynamically expressed among each tissue and contribute to mechanisms that regulate gene expression.

The 3′ Untranslated Region, the portion of the mature mRNA molecule located between the STOP codon and the poly(A)-tail, plays an important role in the regulation of gene expression. These regions of the gene model contain numerous sequence elements that are targeted by non-coding RNAs and RNA-binding proteins that dose gene expression post-transcriptionally [200, 201]. 3′UTRs expression is also dynamically regulated due to a mechanism called alternative polyadenylation (APA), that enables expression of multiple 3′UTR isoforms for the same gene. APA is widespread among eukaryotes [85-89], but the mechanisms that direct APA and its activities in cells of living organisms remain poorly understood.

Cleavage and polyadenylation of eukaryotic mRNAs is directed through the combinatorial action of *cis*-elements near the cleavage site and *trans*-factors within the core polyadenylation machinery [80]. Upstream of the poly(A)-site the hexameric poly(A)-signal element (PAS) is bound by a large multimeric complex known as cleavage and polyadenylation specific factor (CPSF) [77]. This complex interacts with a second large complex called cleavage stimulation factor (CstF), which binds G/U-rich elements downstream of the cleavage site [79].

Together, CPSF and CstF recruit endonucleases to cleave the transcript and poly(A)-polymerase to add the poly(A)-tail [80].

Aside from a few examples [90] [93], there are no unified mechanistic rules described for how the cleavage and polyadenylation machinery discriminates between multiple polyadenylation sites (PAS) in the same 3′UTR. Recent transcriptome-wide 3′UTR mapping efforts in *C. elegans* discovered APA among ~46% of worm genes and have delivered many insights into the mechanisms that direct APA [88, 89]. These analyses uncovered a surprisingly small portion of PAS sites (~39%) enriched with the canonical poly(A)-signal element, 'AAUAAA'. Interestingly, in genes with APA, non-canonical PAS elements that vary by at least one nucleotide from the canonical PAS were more prevalent among cleavage sites proximal to the STOP codon, whereas the canonical hexamer is used predominately for distal cleavage sites [88]. It is still not understood how the cleavage and polyadenylation machinery chooses between different PAS elements in these cases.

We have recently reported that APA is pervasive among *C. elegans* intestine, pharynx, and body muscle tissues, where ~70% of genes among them are expressed with tissue-specific 3′UTRs, suggesting that much of the 3′UTR heterogeneity observed in the worm transcriptome is restricted to specific tissues [128]. Notably, genes expressed with intestine or muscle-specific 3′UTRs are significantly enriched with predicted and experimentally validated miRNA targets suggesting crosstalk between APA and miRNA-induced post-transcriptional gene regulation. These data point to a large potential for genes using APA to selectively exclude or maintain miRNA targets on a tissue-specific basis to precisely dose their expression among them.

So far, this idea has only been genetically validated for a single gene.
In murine quiescent satellite cells, the differentiation-inducing transcription factor Pax3 is maintained at low levels due to the broad expression of miR-166, a miRNA that targets the distal portion of its 3′UTR isoform. In a subset of QSCs, APA of Pax3 enables miR-166 target exclusion, allowing it to counteract miRNA repression and driving its expression to levels sufficient to induce muscle cell differentiation. The widespread nature of alternative polyadenylation in worms coupled with its dynamic expression among tissues and enrichment of miRNA targets suggests this mechanism may play-out in many more cases than what is currently appreciated.

Here, we have applied the PAT-Seq approach to isolate and sequence mRNA from *C. elegans* hypodermis, seam cells, arcade cells, N-methyl-D-aspartate (NMDA) neurons, and gamma-aminobutyric acid (GABA) neuronal cells. We have mapped each of the resulting tissue transcriptomes and remapped our former PAT-Seq derived muscle and intestine transcriptomes [128] to the latest *C. elegans* genome annotation (WS250), which allowed identification of thousands of genes expressed in each tissue. We studied the now refined pools of tissue-specific genes and identified hundreds of genes uniquely important for their functions. Mapping of poly(A)-sites revealed widespread alternative polyadenylation in each tissue transcriptome that allows ~20% of predicted miRNA targets to be lost. We find that, on average, ubiquitously expressed genes among all tissues are longer and more enriched with predicted miRNA targets than tissue-restricted genes suggesting APA as a mechanism to allow precise dosing of these genes between tissues. Ubiquitously expressed genes lose ~40% of predicted miRNA targets to APA events among all tissues and tend to escape distinct miRNAs from tissue-restricted gene sets. Finally, many of the predicted miRNA targets in the proximal portion of the 3′UTRs of the same genes are brought closer to the poly(A)-tail suggesting a dual role for APA in allowing transcripts to either escape from or potentiate miRNA activities.

PAT-Seq from C. elegans hypodermis, seam cells, arcade and intestinal valve cells, and GABAergic and NMDA neurons

We have applied the PAT-Seq approach to isolate, sequence and map mRNA transcripts from *C. elegans* hypodermis, seam cells, GABAergic neurons, NMDA neurons, and cells of the pharyngeal epithelium (**Figure 4.1**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.1. *C. elegans* somatic tissues we have profiled using PAT-Seq. Color-coded boxes correspond to each illustrated tissue.

These cells cover much of the somatic tissue anatomy in worms, allowing for a more comprehensive sampling of tissue-specific transcriptomes to study gene expression and APA dynamics. To further enrich the analysis of tissue-specific transcriptomes in our current study, we have also incorporated datasets from our previously mapped intestine, pharynx, and body muscle transcriptomes [128].

In PAT-Seq, the worm ortholog for the cytoplasmic polyA-binding protein (PABPC), encoded by *pab-1*, is cloned in-frame with GFP and a 3xFLAG epitope, together forming a construct we call PolyA-Pull. PolyA-Pull is then expressed in each tissue of interest using tissue-specific promoters, followed by crosslinking and immunoprecipitation of bound tissue-specific mRNA (**Figure 4.2**). Each of the resulting transcriptome datasets has been incorporated into the latest version of our previously reported APAome.org database [128].

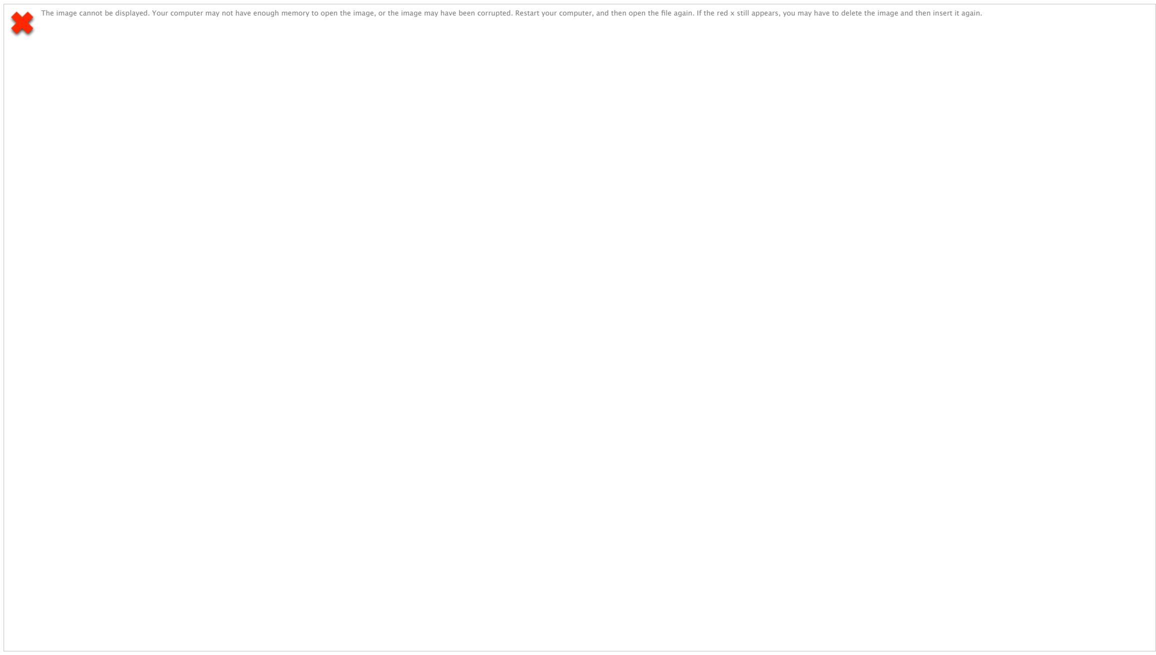
The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.2. Overview of the PAT-Seq method. The Poly(A)-Pull cassette containing GFP fused to 3xFLAG-tagged *pab-1* is fused to tissue-specific promoters. Transgenic worm lines that express this construct are then prepared, grown in liquid culture, followed by crosslinking, lysis and immunoprecipitation of 3XFLAG-tagged PAB-1 complexes. The tissue-specific mRNA extracted from the IP is then used to prepare cDNA libraries for next generation sequencing and transcriptome mapping, which is stored in the APAome.org database.

To ensure efficient immunoprecipitation of mRNA from the smaller tissues, such as the seam cells, we adjusted the PAT-Seq method to make it more sensitive.

In our original protocol, we used the Mos-1 single copy insertion technology to prepare stable transgenic worm lines with a genome integrated polyA-pull cassette [131] [128]. While this method guarantees homogenous expression of the transgene, the low expression levels gained from single copy inserted transgenes may not allow PolyA-Pull expression at levels sufficient for RNA pulldown in tissues composed of just a few cells. We therefore prepared transgenic worm lines expressing PolyA-Pull from multicopy extrachromosomal arrays, which typically overexpress transgenes in somatic tissues[202]. We have also implemented a sonication step to improve worm lysis following crosslinking (**see Experimental**).

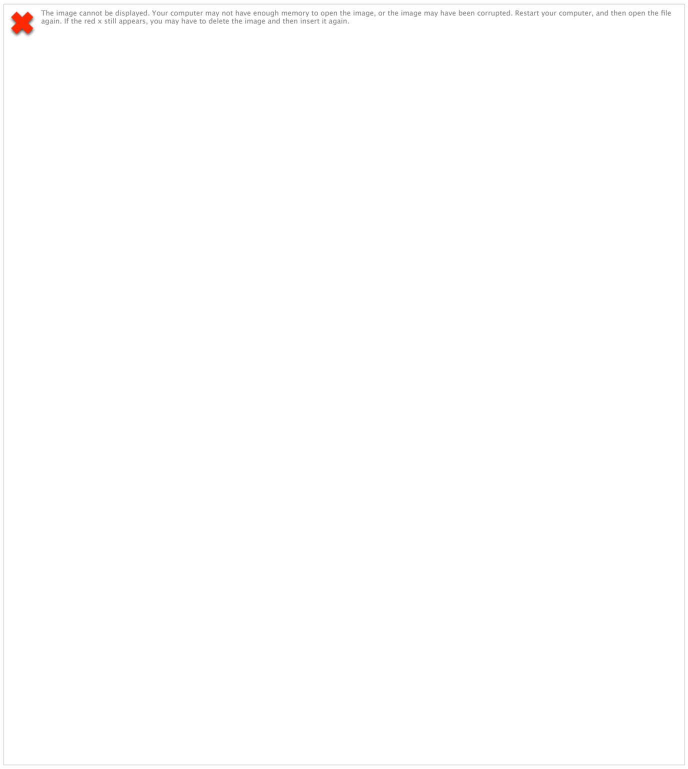
We have prepared transgenic worm lines using tissue-specific promoters to drive expression of the Poly(A)-Pull construct in hypodermis (*dpy-7* promoter), seam cells (*grd-10* promoter), GABAergic neurons (*unc-47* promoter), NMDA-type receptor neurons (*nmr-1* promoter), and the arcade and intestinal valve tissue (*bath-15* promoter) (**Figure 4.3**).



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.3. Fluorescent images of five tissues profiled using PAT-Seq in this study. Representative fluorescent images are of worms expressing the Poly(A)-Pull cassette in each tissue using the indicated promoters. Yellow arrows mark small cells expressing the construct.

After crosslinking and immunoprecipitation of tissue-specific mRNA we used an RT-PCR approach to confirm the enrichment of tissue-specific mRNA. The *dpy-7* and *grd-10* transcripts were selectively enriched in the hypodermis and seam cell mRNA preparations, respectively, while pharynx muscle (*myo-2*) and neuronal cell (*unc-47* and *nmr-1*) transcripts were depleted (**Figure 4.4**). These data indicate that our updated PAT-Seq protocol is sufficiently sensitive to enrich tissue-specific transcripts and specific enough to limit mRNA background from other tissues.



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.4. RT-PCR experiments demonstrating the specificity of mRNA pulldown from hypodermis and seam cells. We detected *dpy-7*, *myo-2*, and *unc-47* transcripts in total RNA from all tissues, while *dpy-7* is specifically enriched in mRNA prepared from *dpy-7*::PAP worms. The same transcripts were not detected in mRNA immunoprecipitated using our negative control construct lacking PABPC (*myo-2*ΔPABP). We specifically detected *grd-10*, but not *unc-47* or *nmr-1* transcripts, from mRNA immunoprecipitated from worms expressing Poly(A)-Pull in the seam cells.

We then prepared cDNA libraries from two biological replicate mRNA pull-down samples for each tissue. As in our former application of PAT-Seq, we used the Single Primer Isothermal Amplification (SPIA) cDNA library preparation technology as it enriches cDNA from small amounts of mRNA and limits mispriming artifacts to improve transcriptome mapping quality [132] [128] (**see Experimental**). We pooled and barcoded the cDNA libraries (10 total) and sequenced them using a 1x50 flow cell on the Illumina Hi-Seq Instrument. This approach yielded millions of reads per sample (**Table 4.1**).

✖ The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Table 4.1. Summary of results from PAT-Seq after deep sequencing.
Raw reads derived from tissue-specific mRNA libraries on the Illumina Hi-Seq Instrument, mapped to the *C. elegans* WS250 genome annotation. Remapped data from our former sequencing is indicated in red text.

In our last application of PAT-Seq, we mapped reads onto the *C. elegans* WS190 worm genome annotation as this version is compatible with miRNA-target prediction datasets, the *C. elegans* 3′UTRomes, and other useful features of the gene model. In recent years, the *C. elegans* genome annotation has been updated with considerable improvements including updated gene name assignments, refining gene coordinates, gene descriptions, and other useful features. We therefore improved our transcriptome mapping for the current study using the latest WS250 worm genome annotation (**see Experimental**). We have also remapped our intestine, pharynx, and body muscle datasets to the latest *C. elegans* WS250 genome annotation, allowing us to integrate these data into our current analysis and improve their quality (**red text in Table 4.1**). Our strategy successfully mapped over 50% of the raw reads from most of the tissue samples (**Table 4.1**). Unmapped reads typically contained stretches of homopolymeric nucleotides, in many cases corresponding to the poly(A)-tail, rendering them ambiguous (*data not shown*).

We were able to map reads corresponding to a similar number of genes between each biological replicate (**Table 4.2**) suggesting that despite the sometimes low portion of reads mapped for a few replicates (**Table 4.1**), this did not interfere with the overall consistency of our approach. The mapped genes and their expression levels between tissue biological replicates also correlated well, further highlighting this consistency (**Figures 4.5 and 4.6**). Although we were unable to map large portions of reads from one of the biological replicates from the seam cells and correspondingly mapped fewer genes in this replicate (**Tables 4.1 and 4.2**), the overall density of genes mapped per their expression value (fpkm) are well correlated (**Figures 4.5 and 4.6**). This data indicates that these replicates are consistent despite the difference in numbers of reads mapped between them.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Table 4.2. Summary of sequencing results after mapping genes to WS250. Mapped reads from the tissue-specific mRNA libraries on the Illumina Hi-Seq Instrument. Genes are mapped to the *C. elegans* WS250 genome annotation. Genes marked with an asterisk correspond to genes enriched in both biological duplicates (fpkm>=1). Remapped data from our former sequencing is indicated in red text.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.5. Correlation between sequenced biological replicate mRNA libraries prepared from hypodermis, seam cells, and NMDA-type neurons. *Left:* Scatter plot of expression values (log10 fpkm) between biological replicates for each gene. *Right:* Histogram plot of the density of genes from each biological replicate falling into expression value bins (log10 fpkm), where red and blue represent each biological replicate.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.6. Correlation between sequenced biological replicate mRNA libraries prepared from GABAergic neurons and arcade and intestinal valve cells. *Left:* Scatter plot of expression values (log10 fpkm) between biological replicates for each gene. *Right:* Histogram plot of the density of genes from each biological replicate falling into expression value bins (log10 fpkm), where red and blue represent each biological replicate.

Epithelial, neuronal, muscle and epidermal tissues display dynamic gene expression profiles

In total, our analysis has assigned almost 11,500 genes to the tissues where they are expressed (**Figure 4.7**). We have profiled transcriptomes that comprise the muscle (pharynx, body muscle), neuronal (GABAergic, NMDA neurons), epithelial (intestine, arcade and intestinal valve cells) and epidermal (hypodermis, seam cells) tissue types and studied their gene expression dynamics (**Figures 4.7 through 4.11**).

The epithelial tissues express the majority of genes detected in our analysis (8,885 genes in total) and express 2,199 genes uniquely (**Figures 4.7 and 4.10**). As previously reported by us and others, this pool contains many genes encoding metabolic enzymes, components supporting innate immunity, and the GATA transcription factors [119, 124, 128, 136]. Interestingly, we detected 73 F-box domain protein genes in this pool, which are the target of ubiquitin ligase complexes that regulate protein homeostasis [203, 204]. These genes are known to participate in epithelial-to-mesenchymal transitions, are commonly misregulated in cancers [205], and potentially regulate the aging process [206].

We detected a much smaller pool (616 genes) exclusively in the GABAergic and NMDA neuronal cells (**Figures 4.7 and 4.9**). As expected, this pool was enriched in common neural genes such as neurotransmitter receptors (cholinergic, FMRF, nicotinic, GABA, and others), calcium and potassium channels, and factors that support axon development.

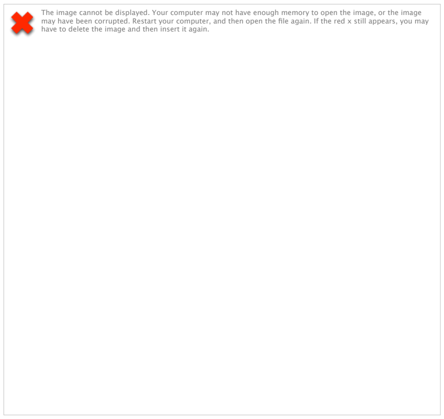
After remapping our pharynx and body muscle datasets to the WS250 genome annotation and subtracting genes expressed in the six tissues profiled in this analysis, we found 360 genes uniquely expressed in these muscle tissues (**Figures 4.7 and 4.8**). Neurotransmitter receptor genes such as dopaminergic, cholinergic, nicotinic, and others were detected in the muscle cells that may function as post-synaptic sites at neuromuscular junctions [207]. These neurotransmitter genes are distinct from those detected in the neuronal datasets. We also found many lectin genes, which support locomotion [144], uniquely expressed in muscles. After subtracting genes expressed in all other tissues, we have now refined our pharynx muscle-specific gene pool to 78 genes uniquely expressed in this tissue (**Figure 4.8**). The body muscle-specific gene pool was refined to 269 genes after subtracting all other tissues (**Figure 4.8**). Similar to what we previously reported, we detected a small overlap of only 13 genes expressed in common between pharynx and body muscle (**Figure 4.8**). Three of these shared genes (*zip-8*, *bed-1* and *klu-2*) are putative transcription factors that we speculate may be important for conferring basic muscle identity.

We detected 920 genes uniquely expressed in the epidermal tissues (**Figures 4.7 and 4.11**). Among them included genes bearing wormbase descriptions that point to an array of roles in molting or embryonic development. Interestingly, we commonly detected several neurotransmitter receptors in the epidermal transcriptome, highlighting the role of the epidermal cells in sensing environment. Consistent with this, serpentine receptor genes, which are important for *C. elegans* chemosensory mechanisms [208], are also abundant in the epidermal gene pool.

Many of the genes uniquely expressed in epithelial, neuronal, muscle, and epidermal tissues do not yet have annotated functions and need to be further investigated.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.7. Venn diagram displaying the portions of genes expressed between four tissue groups we have profiled using PAT-Seq. Tissues were grouped by muscle (pharynx and body muscle), neuronal (GABAergic and NMDA-type neurons), epithelial (intestine and arcade and intestinal valve), and epidermal (hypodermis and seam cells) tissue groups.



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.8. Venn diagram of muscle-unique genes. Illustrated are the portions of genes shared between body muscle and pharynx tissues. We detected a total of 360 genes uniquely expressed in muscle tissues.

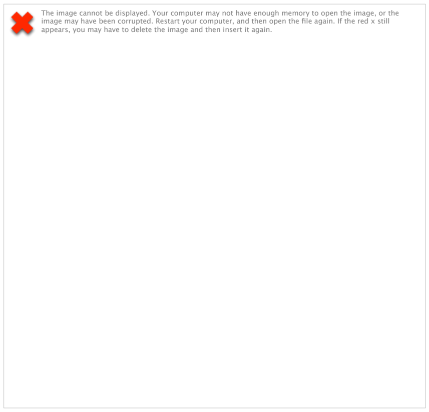

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.9. Venn diagram of neuronal-unique genes. Illustrated are the portions of genes shared between GABAergic and NMDA-type neuronal tissues. We detected a total of 616 genes uniquely expressed in the two neuronal tissues.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.10. Venn diagram of epithelial-unique genes. Illustrated are the portions of genes shared between intestine and arcade and intestinal valve tissues. We detected a total of 2,199 genes uniquely expressed in epithelial tissues.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.11. Venn diagram of epidermal-unique genes. Illustrated are the portions of genes shared between hypodermis and seam cells tissues. We detected a total of 920 genes uniquely expressed in epidermal tissues.

Only 9% of genes in the GABAergic and NMDA-type neuronal transcriptomes are co-expressed

C. elegans adult hermaphrodites have 26 neurons expressing the GABA receptor [209]. 19 of these are D-type motor neurons that span the ventral chord and function to inhibit contraction of the body muscles on one side, opposing the excitatory contraction on the other side to coordinate locomotion [210]. The 5 AME-GABAergic neurons control flexing movements of the head that support foraging activities. Two GABA-neurons, AVL and DVB, are located on the anterior and posterior ends of the worm, respectively and stimulate enteric muscles used for defecation. The remaining GABA-neuron, RIS, is an interneuron with no physiological role defined so far. Cinar et al. have profiled the transcriptome of GABAergic neurons using the mRNA-tagging approach coupled with a microarray detection approach that detected over 250 genes expressed in this tissue [211].

Our NextGen sequencing approach has identified a total of 4,885 genes expressed in this tissue, many of which overlap extensively with the top expressed genes detected using microarray approaches (**Figure 2.12**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.12. Comparison with Cinar et al. mRNA-tagging derived datasets. We compared our GABAergic neuron tissue transcriptome with the microarray-derived GABAergic neuron transcriptome from Cinar et al. (n= 242). *Left:* Venn diagram showing 44% overlap with our GABAergic neuron dataset. *Right:* Cinar et al. genes ordered by expression level (rank) detected in our dataset. We detect most of the top 100 expressed genes from the Cinar et al. dataset.

Of the genes uniquely expressed in neurons, we detected 286 genes uniquely expressed in the GABAergic neurons (**Figure 4.9**). The top expressed genes in this list include potassium channels and several genes previously found important for locomotion. Interestingly, the third most expressed gene in this list is *lin-38*, which encodes a protein with a predicted RNA-binding domain shown to be important for vulva development [212], but as of yet has no neural function described. Out of the 286 unique genes in GABAergic neurons, eighty-six genes in this list have not yet been functionally characterized.

Neurons expressing the ionotropic glutamate receptor enable rapid excitatory neurotransmission. *C. elegans* have ten such receptors, two of which are specifically gated by NMDA agonists and are expressed in six interneurons (AVA, AVD, AVE, PVC, AVG, and RIM) important for locomotory control [213]. We detected many G-protein signaling components (*dmsr-3*, *dmsr-6, rab-37*, and others) and potassium channels (*twk-16*, *twk-17*, and *twk-39*) consistent with neural activities. Interestingly, we detected eight genes belonging to the nematode-specific peptide families that do not have functions described for this tissue. The NMDA-type neuron pool of genes also included transcription factors such as *tbx-34*, a T-box transcription factor with no described function so far. We also many detected worm orthologs of human disease genes. One of these genes is *ceh-6*, a homolog of human POU3F4 transcription factor commonly mutated in conductive deafness [214] (OMIM: 304400) Another, *mbtr-1*, contains human malignant brain tumor repeats (OMIM: 608802) that when mutated in drosophila lead to malignant transformation of the larval brain [215]. These and other disease associated genes point to opportunities to study the role of these genes in normal neuronal phenotypes in worms.

We detected a total of 616 genes uniquely expressed in both neural tissues, but surprisingly only 55 genes are shared between them (**Figure 4.9**). This lack of similarity in their gene expression profiles may reflect the distinct functional differences between these tissues despite their common neural identity. Among the top hits in this list of shared genes are insulin family genes (*ins-1* and *ins-17*), previously shown to be expressed in neurons through development [216] and G-protein signaling components (*rgs-6*, M04G7.3, *srsx-25*, and others). Notably, one of the transcription factors detected in this list is *mab-9*, which is a T-box transcription factor gene known to be enriched in motor neurons and important for axon guidance [217] [218]. We have also detected other transcription factors, such as *hlh-1* and *nhr-67* that may be broadly important for specifying neural identity. Importantly 13 genes (~24%) of the genes in this list do not have a function described yet and their neuronal roles need to be further characterized.

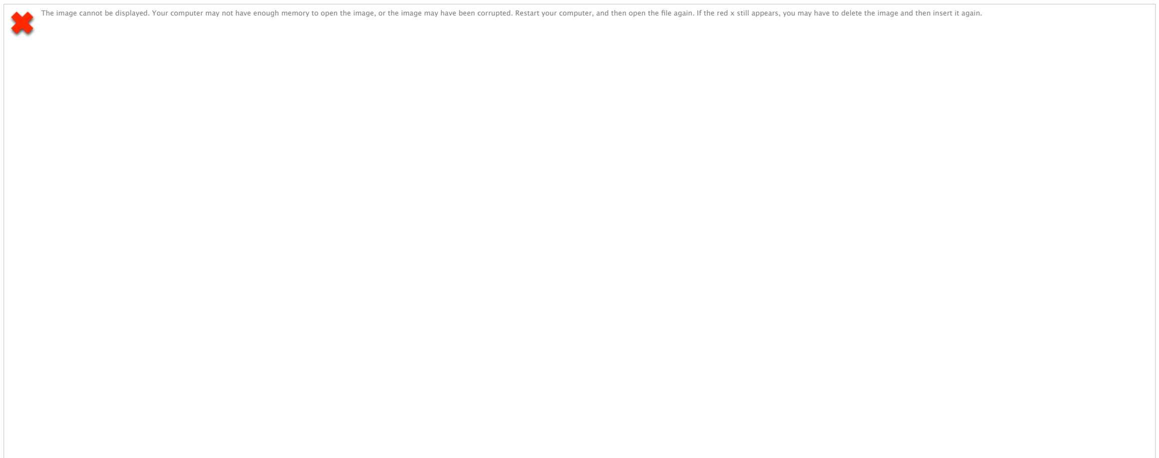
Arcade and intestinal valve cells share large overlap in gene expression with the intestine transcriptome

The epithelium adjacent to the pharynx muscle in the buccal cavity includes arcade cells at the anterior end and the pharyngeal/intestinal valve posterior to the pharynx (**Figures 4.1 and 4.3**). The arcade cells are specialized epithelial cells, organized into two rings, that anchor the pharynx epithelium to the epidermis in the buccal cavity [106]. These cells play an important role in pharyngeal extension during early development of the digestive tract [219]. The six intestinal valve cells posterior to the pharynx muscle form an anatomical barrier between the pharynx and intestine and permit translocation of the physically processed bacterial diet from the pharynx into the intestinal lumen [106]. In the embryo, these cells develop in coordination with the pharyngeal and intestinal primordium to form a wedge-shaped structure that bridges the two organs [220].

We detected a total of 3602 genes expressed in this tissue. This dataset is enriched in genes involved in embryonic morphology and development (*eef-1A.2*, *lev-11*, *icd-1*, and others) and lifespan(dao-6, ril-1, pghm-1, and others). After subtracting genes expressed in all other tissues, we detected 153 genes uniquely expressed in the arcade and intestinal valve cells (**Figure 2.10**). This pool contains ~13 genes with WormBase descriptions indicating a role in embryonic development including *prp-4*, C25E10.11, *clec-233*. There are also several transcription factors with poorly understood functions in this pool that need to be further investigated. We detected a large portion of 211 genes uniquely co-expressed between this tissue and the intestine (**Figure 2.10**), which may reflect the similar roles of intestinal valve cells and the intestinal tract posterior to it. Three of the seven most abundantly expressed genes in this category (*sptf-2*, *nhr-106*, and *elk-2*) are transcription factors that we speculate may play a role in gut formation. Surprisingly, only eleven genes are shared between the pharynx muscle and the arcade and intestinal valve cells (*data not shown*), suggesting at the genetic level this tissue is more similar to its epithelial intestine counterpart.

The hypodermis transcriptome contains over 6,000 genes associated with a vast array of functional activities

The hypodermis is a large epidermal tissue composed of the 138 nuclei hyp7 syncytium, five cells in the head (hyp 1 – hyp5), and four cells in the tail (hyp8 – hyp11) of the adult hermaphrodite [221]. It has multiple functions, including forming the cuticle and basement membranes, directing neuronal placement and influencing axon pathfinding, regulating the development of neighboring cells, removing apoptized cells, and establishing the body plan. We detected a total of 6,033 genes expressed in the hypodermis. Our hypodermis transcriptome overlaps extensively with a published dataset from Spencer et al. derived from mRNA-tagging experiments followed by analysis on tiling arrays (**Figure 2.13**).



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.13. Comparison with Spencer et al. mRNA-tagging derived datasets. We compared our hypodermis tissue transcriptome with the microarray-derived hypodermis transcriptome from Spencer et al. (n= 1,234) *Left:* Venn diagram showing 68% overlap with our hypodermis dataset. *Right:* Spencer et al. genes ordered by expression level (rank) detected in our dataset.

Consistent with its function in cuticle formation, we detected 87 collagen genes expressed in the hypodermis which are genes regulated through development of the animal to coordinate precise molting events [222]. We also detected many hedgehog-related genes, known to be expressed in the hypodermis, and are thought to contribute to cuticle formation in nematodes [223].

We detected 751 genes uniquely expressed in the hypodermis after subtracting genes expressed in the other seven tissues (**Figure 2.11**). Aside from the many cuticle formation genes enriched in this pool, it is also enriched with genes involved in molting (F42A8.1, *fkb-5*, *mlt-10*, *mlt-2*, and others), lifespan and growth rate (*old-1*, *osm-1*, *nphp-4*, F56D5.5 and others), embryonic development (C03B8.2, C17E7.4, C46A5.5, *nphp-4* and others), and solute carriers/transporters (K08H10.6, *snf-12*, T11G6.3, *vglu-2*, *vglu-3*, and others). We also detected transcription factors that are either known to control larval development like *ceh-16* [224] and *nhr-23* [225] or have putative roles in hypodermal cell fate (Y73F8A.24, *ztf-23*, *atf-8*, *attf-5* and others).

Over 31% of the 751 genes (238 genes) we detected uniquely expressed in the hypodermis have no known role described so far.

The seam cell transcriptome contains many epidermal and neuronal genes

Cells located laterally along the hypodermis, called seam cells, undergo a coordinated asymmetric division, forming either hypodermal cells or neural cells, with each molting cycle through larval development [221]. A single symmetric division event at the L2 larval stage produces ten total seam cells that are maintained until the L4 stage, at which point they terminally differentiate into skin cells that produce the adult alae structure [226]. These cells have served as models for stem cell biology as they asymmetric divisions result in a posterior daughter neuronal cell or a hypodermal cell while the anterior seam cell maintains an undifferentiated state [227].

Similar to the hypodermis and consistent with its role in epidermal activities, we detected 52 collagen genes and many molting genes expressed in this tissue. We also detected eight hedgehog-like genes, known to function in this tissue [223]. In contrast with the hypodermis, we detected a much smaller pool of 105 total genes uniquely expressed in the seam cells (**Figure 2.11**). We speculate that these genes expressed in epidermal tissue contribute to unique seam cells identity as opposed that of the hypodermis. Interestingly, this pool of seam cell specific genes contains *ceh-10* and *ceh-43*, which are involved in neural fate specification. Whether these genes may contribute to neural identity specifications of daughter cells resulting from these asymmetric cell divisions during larval development [227] remains to be investigated. Indeed *ceh-43* is expressed in anterior hypodermis and neuronal cells and its loss results in loss of head hypodermal cells culminating in embryonic lethality [228]. Interestingly, a related transcription factor, *ceh-20* was recently shown to be important for controlling seam cell asymmetric divisions [229], providing support for this hypothesis. 67 of the genes uniquely expressed in seam cells have no function described so far and need to be further investigated.

Mapping poly(A)-clusters in hypodermis, seam cells, GABAergic and NMDA neuron expressed genes

We recently provided evidence for pervasive 3′end heterogeneity between intestine, pharynx, and body muscle mRNAs due to a mechanism called alternative polyadenylation (APA) that permits expression of multiple 3′UTR isoforms for the same gene [128]. Our results showed that tissue-specific APA was common and that genes expressed with tissue-specific 3′UTR isoforms are significantly enriched in predicted and experimentally validated miRNA targets [128]. We sought to expand these datasets of tissue-specific 3′UTR expression into the five tissues we have profiled using PAT-Seq and study their expression dynamics.

We employed a strategy similar to that used in our former approach [128] to map poly(A)-clusters and define the 3′ends for the genes detected in our tissue transcriptomes (**see Experimental**). We mapped raw reads containing poly(A) ends from each tissue transcriptome and remapped those from the intestine, pharynx, and body muscle tissues to the WS250 worm genome annotation. To improve the resolution of poly(A)-cluster mapping, we trimmed each of the mapped poly(A) containing reads and scaffolded them onto the WS250 worm genome annotation, then clustered reads for each poly(A)-site. In addition to using Single Primer Isothermal Amplification (SPIA) to limit artifacts that arise from internal priming of oligo-dTs during cDNA library preparation [132] [128], we further removed reads mapping near A-rich stretches of the genome (**see Experimental**).

This approach built poly(A) clusters for ~5,500 hypodermis genes, ~1500 genes expressed in seam cells and arcade and intestinal valve cells, and ~3,000 genes for the GABAergic and NMDA-type neurons (**Table 4.3**). Many of the genes in each tissue express more than one 3′UTR isoform as reflected by the isoforms/gene ratios of each tissue transcriptome (**Table 4.3**).

We also remapped poly(A)-sites using our new strategy for ~6,000 intestine genes, 1,500 pharynx genes, and 1,300 body muscle genes (**Table 4.3**). The number of genes mapped in each tissue is similar to our old approach [128] suggesting that our refined strategy did not compromise mapping quality. We then remapped the *C. elegans* 3′UTRome datasets [88, 89] to the newest WS250 *C. elegans* genome annotation. The majority of the poly(A)-sites mapped in each tissue dataset agrees with both 3′UTRome datasets suggesting that the our refined poly(A) cluster building strategy allowed us to map high-quality 3′ends (**Figure 4.14**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Table 4.3. Mapped poly(A) clusters in each tissue. We have mapped poly(A) clusters using a refined approach to improve resolution of 3′ends that map close to one another. Shown are the numbers of genes and 3′UTR isoforms with mapped or poly(A) clusters. We remapped our intestine and muscle datasets using this new strategy and display the results with red text.

Figure 4.14. Mapped poly(A) clusters are largely supported by the two published *C. elegans* 3′UTRome datasets. We compared the 3′ends mapped by our refined poly(A) cluster mapping approach with those mapped by Mangone et al. or Jan et al. The majority of the poly(A) clusters are supported by either *C. elegans* 3′UTRome dataset. Remapped poly(A) clusters from Blazie et al., 2015 are boxed in red dashes.

Abundant APA within and between tissues induces loss of predicted miRNA targets

We then studied the use of APA in all eight tissues by quantifying the number of genes having more than one 3′UTR isoform in each tissue transcriptome (**Figure 4.15**). With the exception of the hypodermis, each tissue expresses 20-30% genes with multiple 3′UTR isoforms, on average (**Figure 4.15**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.15. Portion of genes expressing multiple 3′UTR isoforms in each tissue. Blue line indicates the portion of genes in each tissue with more than one 3′UTR isoform. Most of the tissues express multiple 3′UTR isoforms for ~10-40% of genes (yellow box), while the hypodermis expresses a much larger portion of genes with APA.

However, a much larger portion of these genes (almost 80% of total genes expressed in each tissue) have multiple poly(A)-sites, indicating that the remaining ~35-40% of genes selectively express tissue-specific 3′UTR isoforms (**Figure 4.16**). These results are consistent with our previous findings that tissue-specific APA is pervasive between intestine and muscles and provides more comprehensive evidence that *C. elegans* somatic tissues use APA in a tissue-specific manner.

Conversely, we have found that the hypodermis instead expresses ~80% of genes with APA, and fewer of them are tissue-specific 3′UTR isoforms (**Figures 4.15 and 4.16**). This large increase in alternative polyadenylation in the same tissue may reflect an increased role for APA during development in this tissue since the hypodermis is known to display dynamic morphology in embryonic development and between larval stages [221].

We have previously shown that genes having tissue-specific 3′UTR isoforms are significantly enriched in predicted and experimentally validated miRNA targets [128], leading to the hypothesis that tissue-specific APA interferes with post-transcriptional regulatory networks driven by miRNAs. This idea has been further supported by results from others who have suggested that miRNA targets lost to APA-induced 3′UTR shortening may allow genes to escape miRNA regulation [111] [96]. We hypothesized that miRNA target loss due to the APA observed in each tissue transcriptome may be dynamically regulated between tissues. To investigate this idea, we downloaded miRNA target prediction data from PicTar [57] and miRanda [230] databases and mapped the predicted targets to the tissue transcriptomes (**see Experimental**). We used a stringent approach to select the most abundant 3′UTR isoform expressed for each gene and quantified the number of predicted miRNA targets lost due to each unique APA event (**Figure 4.16**). On average, each tissue lost ~20% of the predicted miRNA targets due to APA-induced 3′UTR shortening (**Figure 4.16**). However, the NMDA-type neurons lost only ~3% predicted targets, presumably due to a preference for long 3′UTR isoforms and therefore maintenance of miRNA targets (**Figure 4.16**). Conversely, the intestine loses ~33% and the hypodermis transcriptome loses greater than 60% of its predicted miRNA targets indicating a high propensity for genes expressed in these tissues to escape miRNA regulation (**Figure 4.16**).

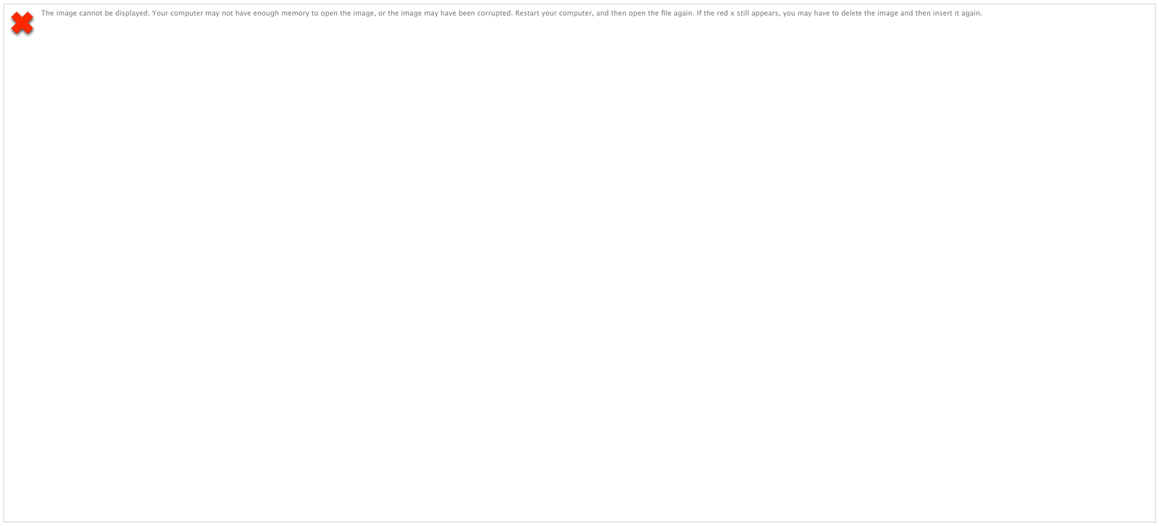
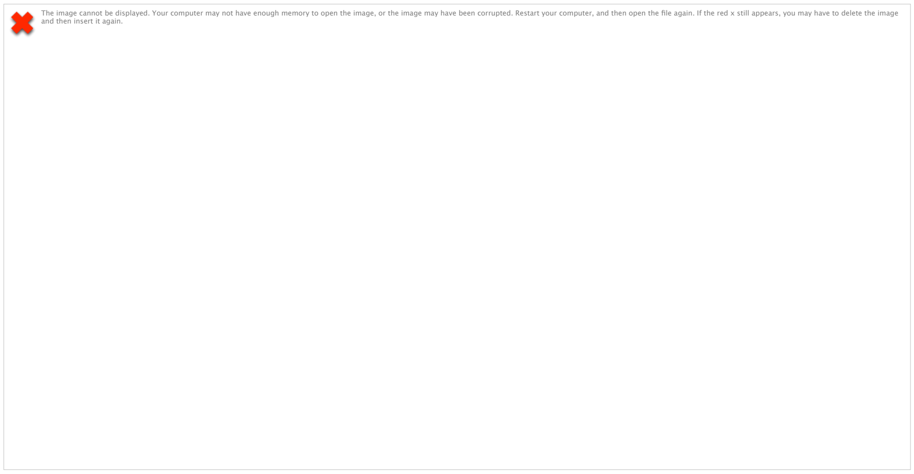

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.16. Dynamics of miRNA target loss driven by APA in each tissue. Displayed are the portions of PicTar or miRanda targets lost between tissue transcriptomes (light purple bars) due to APA. The portion of genes having multiple 3′UTR isoforms within or between tissues is indicated (blue line).

Commonly expressed genes lose distinct predicted miRNA targets due to APA

We then hypothesized that each tissue transcriptome may prefer to escape distinct miRNAs to buffer their gene expression. We focused on the general enrichment of *C. elegans* miRNA families, which have the same 'seed' sequence used to target the mRNA [231] and ranked their general enrichment in genes expressed within each tissue transcriptome regardless of the 3′UTR they expressed (**Figure 4.17**). The top five most enriched miRNA families are strikingly similar between each tissue transcriptome (**Figure 4.17**). In particular, predicted targets of miRNA families *miR-2*, *let-7*, and *miR-58* dominated the top three most enriched miRNAs in every tissue and only six miRNA families (ex. *miR-1*, *miR-86*) appeared in the top five ranks of less than two tissue transcriptomes (**Figure 4.17**). We found that the same miRNA families are also enriched in the pool of genes that are commonly expressed among tissues (**Figure 4.17**).

Next, we ranked the miRNA families with predicted targets that are lost in the same genes due to alternative polyadenylation using the most abundantly expressed 3′UTR isoform (**see Experimental, Figure 4.18**). Surprisingly, the most frequently lost miRNA families (**Figure 4.18**) have distinct identities from those that are simply enriched in genes in each tissue transcriptome (**Figure 4.18**). However, the predicted miRNA targets that are most frequently lost are highly similar between tissues where only four families (*miR-2*, *miR-58*, *miR-72*, and *miR-34*) are lost in fewer than three different tissues (**Figure 4.18**). This data suggests that each tissue transcriptome is frequently escaping many of the same miRNAs. In contrast, four of the top five most frequently lost miRNA family targets among genes that are commonly expressed among all eight tissues have a very different profile of predicted miRNA targets lost (**Figure 4.18**). Notably, miRNA families frequently lost in each tissue transcriptome such as *lin-4* and *let-7* are known to be expressed in many different tissues [232] [233]. Conversely, miRNA families such as miR-1, which is frequently lost among commonly expressed genes, is known to be expressed only in the body muscle [234]. We speculate this may point to a mechanism where tissue-specific genes most frequently escape commonly expressed miRNAs and commonly expressed genes more often escape tissue-specific miRNAs to buffer gene dosage in each unique environment where they are expressed.



The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.17. Enrichment ranking of miRNA family targets predicted in each tissue-transcriptome. We downloaded PicTar and miRanda miRNA target prediction data for each tissue transcriptome and the pool of 777 genes that are expressed in every tissue (commonly expressed). We then grouped the predicted targets by miRNA families. Shown are the top five most abundantly predicted miRNA family targets, where #1 is most enriched. miRNA family names are color-coded. Predicted miRNA family targets uniquely enriched in fewer than three tissues are highlighted in yellow with red boxes. The enrichment of predicted miRNA targets between tissues is similar.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.18. Enrichment ranking of miRNA family targets lost due to APA in each tissue-transcriptome. We downloaded PicTar and miRanda miRNA target prediction data for each tissue transcriptome and the pool of 777 genes that are expressed in every tissue (commonly expressed). For each set of genes, we selected the most abundantly expressed 3′UTR isoform for each gene and counted the predicted miRNA targets (grouped by miRNA families) that are lost. Shown are the top five most abundant predicted miRNA family targets lost to APA, where #1 is most enriched. miRNA family names are color-coded. Predicted miRNA family targets that are lost to APA and uniquely enriched in fewer than three tissues are highlighted in yellow with red boxes. The enrichment of predicted miRNA targets lost to APA between tissues is similar, while those in the commonly expressed gene pool are mostly distinct.

Commonly expressed genes express longer 3′UTRs and 85% of them use alternative polyadenylation

We previously reported that *C. elegans* pharynx and body muscle tissues express slightly longer 3′UTRs, while the intestine expresses 3′UTRs closer to the overall median length of ~140 nucleotides observed in the worm 3′UTRome [128] [88, 89]. Recent transcriptome RNA sequencing efforts have provided evidence for generally longer 3′UTRs expressed in mammalian brain tissue that are enriched with neural miRNA targets [98]. We hypothesized that *C. elegans* GABAergic and NMDA-type neurons may also prefer longer 3′UTRs. After remapping poly(A)-clusters for our muscle and intestine datasets and remapping those for the five tissues profiled in this study, we found that median 3′UTR length was similar between tissues with common roles (**Table 4.4**). Muscle tissues express the longest 3′UTRs with medians over 200 nucleotides as previously reported [128]. Both neural tissues express longer 3′UTRs, on average, having 3′UTR median lengths slightly less than 200 nucleotides. Seam cells, arcade and intestinal valve cells, and hypodermis 3′UTRs are slightly shorter, with median lengths averaging ~180 nucleotides (**Table 4.4**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Table 4.4. Median 3′UTR length for all genes expressed in each indicated tissue. Tissues are sorted by ascending median length.

We then extracted tissue-restricted genes, those expressed uniquely in each tissue, and binned them by 3′UTR length (**Figure 4.19**). This revealed a consistent distribution of 3′UTR lengths between tissues that were on average shorter than the *C. elegans* 3′UTRome (colored lines versus dotted line in **Figure 4.19**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.19. Histogram displaying the length distribution of 3′UTRs of tissue-restricted genes. 3′UTRs of tissue-specific genes (colored lines) have very similar lengths and are slightly shorter, on average, than all genes in the 3′UTRome (black dotted line).

Conversely, 3′UTRs of genes expressed in all eight tissues (commonly expressed genes) were on average much longer than tissue-restricted genes and also longer than the 3′UTRome overall (**Figure 4.20**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.20. Histogram comparing the length distribution of 3′UTRs of tissue-restricted genes or commonly expressed genes. Portion of genes in each length bin that are tissue-specific are denoted by squares and commonly expressed genes are denoted by triangles. Commonly expressed genes have longer 3′UTRs, on average.

Commonly expressed genes also express over three times more 3′UTR isoforms per gene than tissue-restricted genes (**Figure 4.21**). Consistent with this observation, 85% of commonly expressed genes use APA between tissues (**Figure 4.22**). Together, these data suggest an increased role for APA in regulating the expression of ubiquitously transcribed mRNAs.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.21. APA in tissue-restricted versus commonly expressed genes. Number of tissue-specific or commonly expressed genes (dark gray bars) and their 3′UTR isoforms (light gray bars) are displayed and designated by the left axis. The isoforms per gene ratio in each pool of genes is displayed by the blue line and designated by the right axis. Commonly expressed genes have many more 3′UTR isoforms per gene, on average.

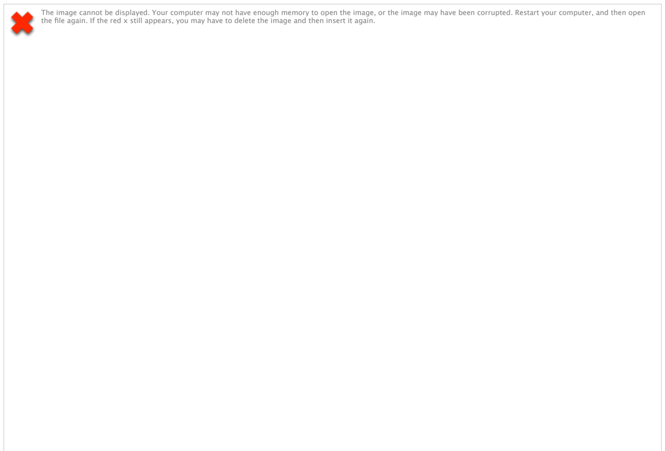

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.22. Portion of genes with APA in tissue-restricted versus commonly expressed genes. Pie charts display the proportion of tissue-restricted genes (left chart) or commonly expressed genes (right chart) with more than one 3′UTR isoform (genes with APA). Nearly all commonly expressed genes are prepared with more than one 3′UTR isoform.

Cleavage and polyadenylation is guided by the poly(A)-signal (PAS) element, a hexamer positioned ~19 nucleotides upstream of the cleavage site [80]. *C. elegans* uses the canonical PAS element ‘AAUAAA’ for only 39% of cleavage events and variant PAS elements are frequently used, especially in genes that use APA [88, 89]. We mapped PAS sites using the iterative procedure previously described [128] (see Experimental). We found that tissue-restricted genes use the canonical PAS at a similar rate to that observed in the 3′UTRome (**Figure 4.23**). NMDA-type neurons are the large exception and appear to use the canonical PAS element more frequently (~50% of NMDA-type neuron restricted genes) (**Figure 4.24**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.23. PAS usage in tissue-restricted versus commonly expressed genes. Pie charts display the proportion of 3′UTR isoforms expressed with tissue-restricted genes (left chart) or commonly expressed genes (right chart) using the canonical PAS AAUAAA or all other PAS elements. 3′UTR isoforms of commonly expressed genes use the canonical PAS slightly less often.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.24. Frequency of PAS element usage in tissue-restricted genes. Portion of 3′UTRs of tissue-restricted genes that use the canonical PAS ‘AAUAAA’ (blue) or other variant PAS elements (red) are indicated. Seam cells and muscle tissues have been omitted from this analysis because too few poly(A) clusters were mapped for their tissue-restricted genes. Genes uniquely expressed in NMDA neurons use canonical and variant PAS elements at a similar rate.

In contrast, commonly expressed genes use the canonical PAS elements slightly less frequently than tissue-restricted genes (26% compared to 40%) and more often use variant PAS elements (74% compared to 60%) (**Figure 4.23**). This increased usage of the variant PAS elements in commonly expressed genes likely reflects their more abundant use of APA and points to a role for variant PAS elements in controlling tissue-specific APA events. We further studied PAS usage for the pool of commonly expressed genes in each tissue individually (**Figure 4.25**). Interestingly, the hypodermis and intestine use variant PAS elements more often than each of the other tissues, while GABAergic and NMDA-type neurons use canonical PAS elements more frequently (**Figure 4.25**). We believe this preference for variant PAS elements in the intestine and hypodermis reflects their general increased loss of miRNA targets through APA-induced 3′UTR shortening events (**Figure 4.25**).

Similarly, the preference for canonical PAS elements in neural tissues likely reflects their preference for APA-induced longer 3′UTRs and, in the NMDA-type neurons, lower rates of miRNA target loss (**Figure 4.16**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.25. Frequency of PAS element usage in commonly expressed genes. Portion of 3′UTRs of commonly expressed genes that use the canonical PAS ‘AAUAAA’ (blue) or other variant PAS elements (red) in each tissue are indicated. The neuronal tissues use canonical and variant PAS elements at a similar rate.

Alternative polyadenylation modulates the presence and positioning of predicted miRNA targets in 3′UTRs of commonly expressed genes

Since commonly expressed genes appear to use APA more frequently than tissue-restricted genes and we have previously shown that genes expressing tissue-specific 3′UTR isoforms are enriched in miRNA targets [128], we hypothesized that commonly expressed genes are also enriched in miRNA targets. We observed predicted miRNA targets in 75% of genes commonly expressed between tissues (**Figure 4.22**).

This frequency is much greater than tissue-restricted genes, of which only 35% are predicted to be targeted by miRNAs (**Figure 4.22**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.26. Portion of commonly expressed or tissue-restricted genes with predicted miRNA targets. Commonly expressed genes have more predicted miRNA targets than tissue-restricted genes.

Commonly expressed genes express much longer 3′UTRs than tissue-restricted genes, on average (**Figure 4.26**), and therefore contain more sequence landscape to harbor such targets. However, commonly expressed genes are also more enriched for miRNA targets in 3′UTRs of similar length to 3′UTRs of tissue-restricted genes, including short 3′UTRs that are less than 200 nucleotides in length (**Figure 4.27**).

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.27. Predicted miRNA target density in 3′UTRs of tissue-restricted versus commonly expressed genes. Commonly expressed genes (squares) of a similar length harbor more predicted miRNA targets than tissue-restricted genes (triangles).

These data suggest that commonly expressed genes have greater potential for post-transcriptional regulation by miRNAs. We therefore hypothesized that the extensive APA observed in commonly expressed genes (**Figure 4.21**) coupled with the abundance of predicted miRNA targets could potentiate miRNA target loss. We selected the shortest 3′UTR expressed for each gene and quantified the predicted targets loss due to these specific APA events (**Figure 4.28**). This analysis shows that ~43% of predicted miRNA targets in commonly expressed genes are lost due to APA (**Figure 4.28**), suggesting a potent role for APA in modulating the activities of miRNAs in *C. elegans* somatic tissues.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.28. Predicted miRNA targets in commonly expressed genes lost to APA-induced 3′UTR shortening. We quantified predicted miRNA targets that are lost in all APA events where the shortest 3′UTR isoform is expressed in commonly expressed genes. APA affects almost half of all predicted miRNA targets in commonly expressed genes.

Although APA appears to relieve commonly expressed genes from 43% predicted miRNA targeting events, it is not clear if APA events may play a role in modulating activities of the 57% of targets that remain after 3′UTR shortening. A recent study provided biochemical evidence that mRNA transcript deadenylation is enhanced when miRNA targeting occurs proximal to the poly(A)-tails of mature mRNAs [235]. We therefore hypothesized that APA-induced 3′UTR shortening events could potentiate miRNA activities by bringing the remaining 57% targets closer to the poly(A)-tail. We binned miRNA target position relative to their distance from the STOP codon for each of three categories, 1) genes without APA expressed with a single 3′UTR isoform, 2) commonly expressed genes expressed with the longest 3′UTR isoform, and 3) commonly expressed genes expressing the shortest 3′UTR isoform (**Figure 4.29**). Predicted miRNA targets in 3′UTRs of genes without APA show a relatively even dispersion throughout the length of the 3′UTR, with a slight preference for the distal end near the poly(A)-tail (orange dashed-box in **Figure 4.29**). In commonly expressed genes expressing the longest 3′UTR isoform, the miRNA targets that are not lost due to shortening are mostly located close to the STOP codon (red dashed-box in **Figure 4.29**). Due to APA, shortening of the 3′UTR for the same pool of genes results in a shift of the poly(A)-tail position proximal to the remaining miRNA targets (**Figure 4.29**), potentially allowing increased activities of the corresponding miRNAs. These results highlight a role for APA in driving not just miRNA target loss, but also potentiation of miRNA activities by controlling the proximity of targeting events to the poly(A)-tail.

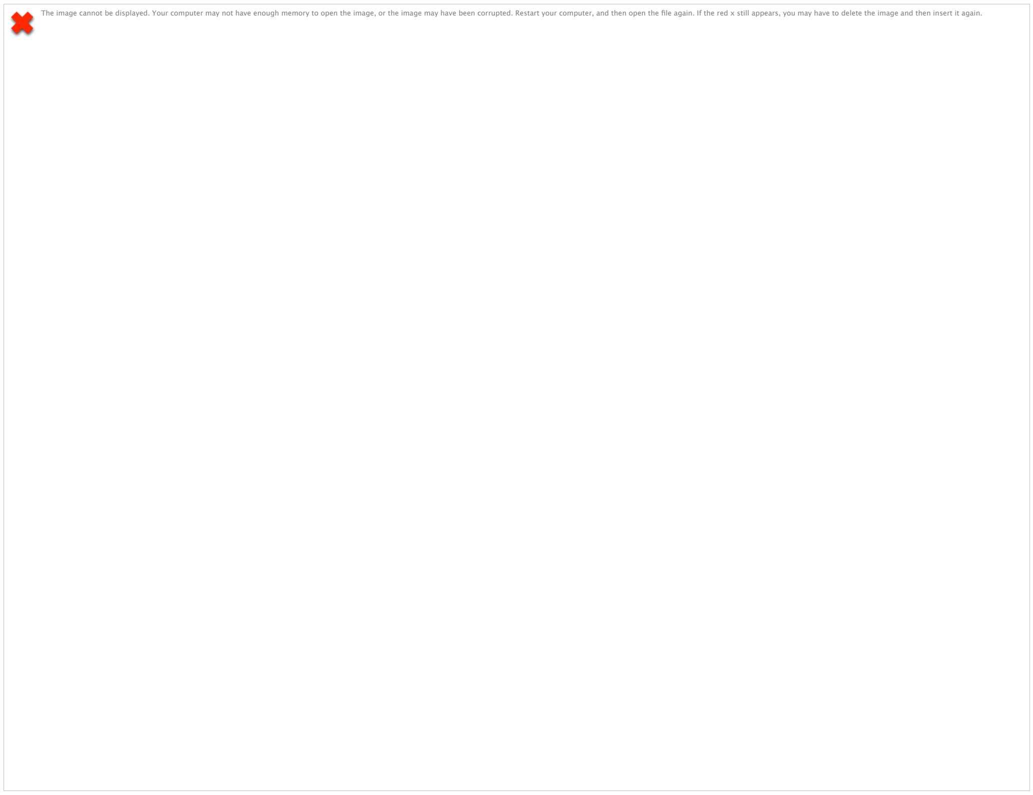

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 4.29. Repositioning of predicted miRNA targets due to APA. We binned predicted miRNA targets by their relative distance from the STOP codon and studied the change in position of these targets relative to the poly(A)-tail after APA-induced 3′UTR shortening. Top: Portion of predicted miRNA targets that remain in commonly expressed genes after APA-induced 3′UTR shortening, shown as purple peaks. When the long 3′UTR is expressed through usage of the distal PAS, the same predicted miRNA targets are located centrally. Conversely, predicted miRNA targets are brought closer to the poly(A)-tail when the proximal PAS is used. Bottom: In genes expressed with a single 3′UTR isoform (n=114), predicted miRNA targets are positioned more evenly along the 3′UTR.

Discussion

PAT-Seq enables isolation and sequencing of mRNA from small worm tissues

The mRNA-tagging method, developed almost fifteen years ago, was initially used to profile mRNA from large tissues for microarray analysis. Since then, several groups have improved it to allow isolation of mRNA from smaller tissues such as neurons [126, 130, 236]. We have recently coupled the mRNA-tagging with NextGen sequencing [237], making it useful for sequencing of tissue-specific mRNA and mapping their mRNA isoforms. However, it was unclear if this approach could be used to profile mRNA from small tissues composed of just a few cells.

Here, we have used PAT-Seq to profile mRNA from tissues either composed of just a few (arcade and intestinal valve and NMDA-type neurons) or having smaller (GABAergic neurons and seam cells) cells. Using an RT-PCR approach, we have shown that the RNA pull-down is specific enough to enrich for tissue-specific transcripts from the seam cells with little background from surrounding tissues (**Figure 4.4**), consistent with others who have performed RNA pull-downs from small neurons [130, 211, 236]. Sequencing the mRNA from these pull-downs enabled us to detect significantly more genes with increased sensitivity compared to past approaches. This is reflected by our sequencing results from GABAergic neurons, which correlate with the top expressed genes from other groups and significantly expand the number of genes detected in this tissue from just over 200 genes to almost 5,000 genes. Our results highlight PAT-Seq as a sensitive and specific method for profiling small tissue transcriptomes in *C. elegans*.

Refined pools of tissue-specific expressed genes reveal hundreds of genes that contribute to their specialized functions

Our application of PAT-Seq to profile three large tissues from our former study [128] and five additional tissues reported here has now assigned ~60% (almost 12,000 genes) of the protein-coding genome to tissues where they are expressed. Our tissues span the four major cell-type categories in worms: muscle, epithelial, neuronal, and epidermal. We have studied the dynamics of gene expression between these tissue groups and found many genes unique to each group that we speculate give rise to their identities. We also found tissue-specific transcription factors that likely contribute to the transcriptional programs conferring identity to each.

The comprehensive nature of our approach has allowed us to further refine the pools of genes expressed uniquely in each tissue pool. Other tissue specific RNA-Seq approaches in *C. elegans*, including our recent application of the PAT-Seq method, typically restricts analysis to just a few tissue types at a time. These approaches are limited in that they cannot properly study pools of tissue-specific genes relative to many other somatic tissues as we report here. Our former study of worm intestine and muscles uncovered hundreds of genes specific to muscle tissues and over four thousand in the intestine. We have now refined these pools to only 78 pharynx-specific genes, 269 genes in the body muscle, and 1,643 specific to the larger intestine tissue. The genes we find expressed in these tissues match their specialized physiological roles and will allow for a more close examination of the genes specifically required for their functions.

Dynamic loss of predicted miRNA targets to 3′UTR shortening

We have shown that ubiquitously expressed worm genes *rack-1* and *tct-1* use APA to escape miRNA regulation by *miR-50*, a miRNA that is also widely expressed among worm tissues. Although these experiments limited the validation of APA as a mechanism that genes can escape miRNA regulation to just two genes, the widespread nature of APA in *C. elegans* and its correlation with miRNA enrichment suggests this mechanism is operating on a much wider level that what is currently appreciated. This provoked our more comprehensive analysis of the effect of APA on predicted miRNA regulatory networks operating among worm transcriptomes reported here. Using target predictions, we find that APA appears to be driving target loss at a rate of ~20% in each tissue-transcriptome, on average. Interestingly, APA-induced miRNA target loss was greatest in the hypodermis and intestine, which are two tissues known to be involved with a large number of physiological roles. On the other hand, target loss due to APA was minimal in the NMDA-type neurons, which are more specialized tissues. These data lead us to speculate that the extent of 3′UTR shortening is correlated with the functional capacity of tissues. Hypothetically, many more genes may need to be upregulated in the hypodermis and intestine to support their large array of functions, while fewer functions are needed in the NMDA neurons where APA-induced 3′UTR lengthening and gene repression appears to be more common. More experiments are needed to clarify the specific role of APA in each of these tissues. Others have reported widespread shortening of 3′UTRs due to misregulation of APA in cancer cells [96] [238]. Our results suggest APA may also operate in normal states to selectively remove genes from post-transcriptional regulatory networks driven by miRNAs.

Commonly expressed genes are subject to greater regulation

We have now profiled mRNA from eight worm tissues and mapped their poly(A) sites and studied a refined pool of 777 genes that are transcriptionally expressed in every tissue. We found that these genes have much longer 3′UTRs with greater predicted miRNA target enrichment, and more APA than tissue-restricted genes. These data point to a model where commonly expressed genes use APA to precisely buffer their expression levels between tissues by selectively escaping miRNAs.

Interestingly, Chen and colleagues have recently shown that average 3′UTR length in metazoans is proportional to their number of tissue types [239]. These data suggest that activities driven by sequences in metazoan 3′UTRs may play a role in conferring tissue identity. Our finding that 3′UTRs of commonly expressed genes are substantially longer than the worm 3′UTRome highlights the idea that these genes may require more sequence landscape to support the varying 3′UTR mediated regulatory activities across each tissue.

Our results are also supported by a recent survey of APA in a consortium of human cell lines by Lianoglou and colleagues, which showed that APA is frequently used as a mode of post-transcriptional regulation among commonly expressed genes [111]. Lianoglou et al. also reported an interesting correlation where genes commonly transcribed among tissues may use APA to counteract repression mediated by ubiquitously expressed miRNAs. Indeed, our examination of the relative enrichment of predicted miRNA targets revealed a distinct profile of miRNAs that are lost in commonly expressed genes compared to those lost in individual tissue-transcriptomes suggesting this may be the case in worms. It is not clear, however, whether the expression pattern of these miRNAs is tissue ubiquitous or tissue-restricted. Further experiments that address their specific expression patterns will shed insight on this idea.

Alternative polyadenylation as a mechanism to modulate miRNA targeting activities between tissues

We have found an interesting dynamic between alternative polyadenylation and miRNA target positioning that may indicate a dual-role for tissue-specific APA. Approximately ~85% of commonly expressed genes express more than one 3′UTR isoform between tissues. Among these APA events, 3′UTR shortening drives loss of ~46% of the predicted miRNA targets in these genes. We believe this target loss

We have also found that the 57% targets that remain after APA shortening are brought closer to the poly(A)-tail, which may influence the activities of the miRNAs that target them. In the cases where the distal PAS is used resulting in long 3′UTR expression, the same miRNA targets are otherwise positioned in the central region of the 3′UTR, which is associated with dampened miRNA activities [240]. Two recent studies have shown that 3′UTR shortening induced by APA can position miRNA targets closer to the poly(A)-tail, augmenting their functionality [235, 241] .

Our results are consistent with and complement the results of both studies. We show that this effect is prominent among the pool of genes that are commonly expressed among eight somatic worm tissues. These data suggest a model where APA events that express short 3′UTRs can lead to two results, 1) miRNA target exclusion enabling selective removal from miRNA targeting networks or 2) re-positioning of miRNA targets proximal to the poly(A)-tails increasing their potency.

These mechanisms may not be mutually exclusive and highlight the context-dependent nature of APA among somatic tissues in *C. elegans*.

Experimental

Plasmids and molecular cloning

Molecular cloning of the PolyA-Pull and Δ*pab-1*-pull plasmids have been described previously [237], and were used in this manuscript with no modifications. The tissue-specific promoters used in this study were selected as up to 2kb of genomic sequence located between the start codon of the target gene (WS250) and stop codon of the next closest gene. The primers were designed using the University of Santa Cruz (UCSC) Genome Browser software with 5-prime Gateway-compatible recombination (Invitrogen) elements for cloning into pDONR P4-P1R entry plasmid (**Table 4.5**). The DNA promoter elements were amplified using PCR from N2 genomic DNA and cloned into Gateway™ pDONR P4-P1R entry plasmids. We used Multisite recombination reactions (LR Clonase II plus, Invitrogen) to combine the tissue-specific promoters with the PolyA-Pull and the *unc-54* 3′UTR into the destination plasmid pCFJ150, which contains the *unc-119* rescue cassette.

Nematode strains and transgenesis

The EG4322 background worm strain, which we used to prepare PolyA-Pull expressing transgenic worm lines, were maintained at 16°C on NGM plates seeded with HB101 bacteria prior to microinjection. Extrachromosomal array transmitting worm lines were prepared by microinjecting pCFJ150 Tissue-specific Promoter::GFP::*pab-1*::*unc-45*

3′UTR (25ng/µl) along with the markers pJL43.1 (50ng/µl), pCFJ90 (1ng/µl), pGH8(10ng/µl), pCFJ104 (5ng/µl) into the background worm strain EG4322 (ttTi5605; *unc-119*(*ed9*) *III*), which were kindly provided by Priscilla Van Wynsberghe (Colgate University, Hamilton, NY, USA). Microinjection was carried out as described previously [168] using a Leica DMI3000B microscope.

RNA immunoprecipitation

Mixed stage cultures of each transgenic worm line were grown in liquid culture at 20°C as described [152]. Worms harvested from liquid culture were crosslinked in formaldehyde, and flash frozen as previously described [237]. Worm lysates were prepared as follows: each worm pellet was thawed, centrifuged for 30 seconds at 10,000 RPM and resuspended in lysis buffer (150 mM NaCl, 25 mM HEPES, pH 7.5, 0.2mM dithiothreitol). Samples were then sonicated on ice 5-times (10 second pulses, 18W RMS output) with 50 seconds pauses between pulses. Lysates were centrifuged at 16,000 x *g* for 15 minutes at 4°C. The supernatant from each lysate was used for immunoprecipitation of mRNA, which was carried out as previously described [237]. Each sample was quantified using the Nanodrop Instrument (Thermo Scientific) and subsequently used in RT-PCR reactions and cDNA library preparation for sequencing.

RT-PCR reactions

For each tissue-specific RNA sample, 100ng RNA was reverse transcribed with an $NVdT_{(23)}$ primer (**Table 4.5**) using Superscript Reverse Transcriptase III (Thermo-Fisher Scientific), according to the manufacturer’s protocol. One microliter of the resulting cDNA was used as a template for each second DNA strain reaction, along with 1μM of gene-specific primer (**Table 4.5**) and Taq DNA Polymerase (NEB) to drive the reactions.

PolyA cluster building and mapping

PolyA clusters were built using custom Perl scripts. We extracted FASTQ sequence reads containing at least 23 adenosines at their 3′ ends, removed the A's and mapped the remaining sequence (≥10nts) to the WS250 worm genome annotation using Bowtie [242]. For each aligned read we selected 5nts upstream and downstream of the sequence region surrounding the 3′ end of each mapped read and built PolyA clusters from overlapping 3′end sequence fragments using BedMerge [243]. We ignored PolyA clusters mapping to regions containing ≥65% adenosines within 20nt of the end of each cluster. Each PolyA cluster was then bioinfomatically attached to the closest WS250 gene on the same strand within no more than 100nt downstream of the furthest WS250 defined 3′UTR end. We merged PolyA clusters mapping within ≤5nt across all the datasets. We ignored PolyA clusters having less than 5% of the total reads for all clusters in a given 3′UTR.

miRNA target prediction analysis

We downloaded *C. elegans* miRNA target prediction data from the PicTar [57] and miRanda [230] databases and obtained the miRNA name and target coordinates for each mapped 3′UTR of each gene. To study the enrichment of predicted miRNA family targets between tissue transcriptomes, miRNA targets were grouped into their families [231] using custom VBA scripts in Microsoft Excel. We ignored miRNAs not belonging to *C. elegans* families [231].

Table 4.5. Primers used in chapter 4.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

CHAPTER 5

CONCLUSION

The genome sequencing era ushered in new insights into how a genome gives rise to so many diverse phenotypes in multicellular organisms. Current Ensembl data estimates the absolute numbers of protein-coding gene loci mapped in the DNA of many common multicellular model organisms is surprisingly similar, despite their obvious differences in morphological complexity [244] (**Figure 5.1**). Therefore, metazoans are much more than simply the sum of their protein-coding parts. Instead, it is becoming clear that the combinatorial variation in gene expression driven by a variety of mechanisms extending from the epigenetic to the post-translational level finely tunes gene expression and establishes cell identity. Past dogma that cellular protein and transcription levels in cells are directly proportional is no longer considered accurate. The discovery of vast regulatory non-coding RNAs in metazoans over the last twenty-five years has established that dosing of gene expression at the post-transcriptional level is a common and potent mechanism contributing to tissue development and maintenance of homeostasis.

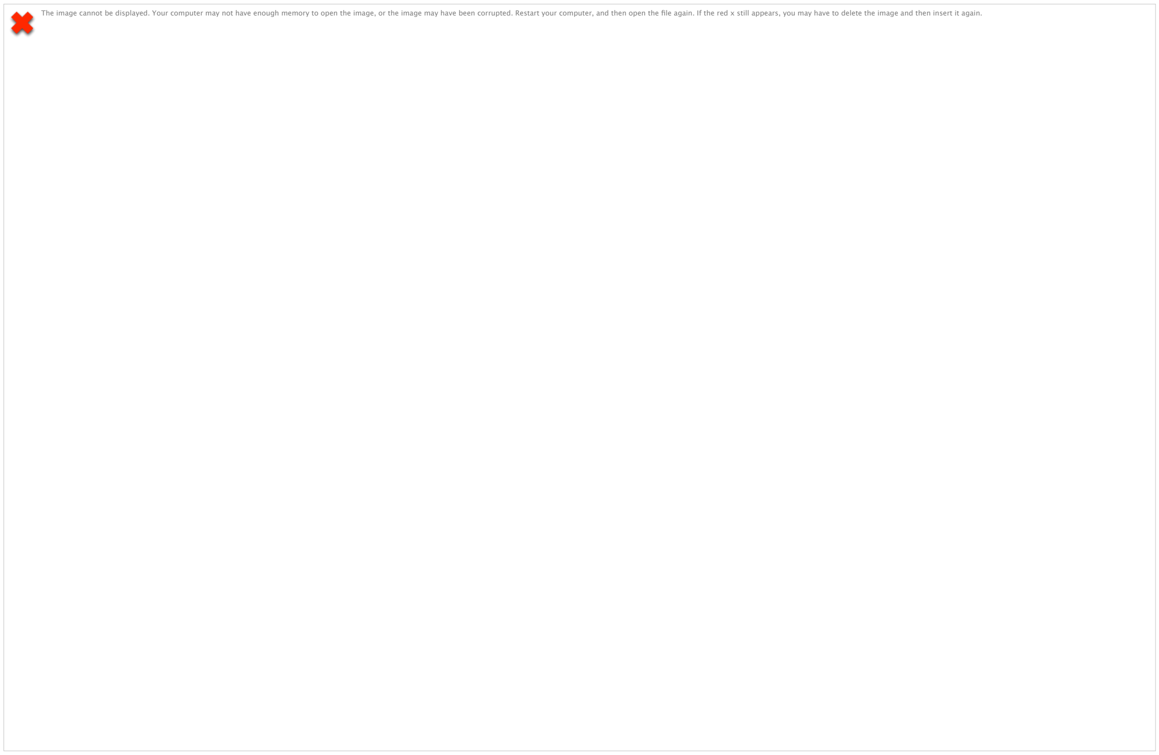

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 5.1. Protein-coding gene count is similar among metazoan genomes. We used data from the Ensemble database [244] of estimated protein-coding genes for the listed organisms and plotted them by absolute number of genes in ascending order. Despite their differences in morphological complexity, the invertebrates and vertebrates only show a modest increase in number of protein-coding genes.

The expression of multiple 3′UTR isoforms for the same gene through APA has long been suspected as a mechanism that substantially contributes to these programs by controlling the variation of regulatory target elements in the 3′UTR landscape. However, its complex nature has made it challenging to detect general rules for its mechanisms and activities in mammalian models where comprehensive 3′UTR annotations are generally not available. The recent exhaustive sequencing of *C. elegans* 3′UTRs, coupled with its genetic tractability, makes this organism a facet to detect general rules for this process that can then be applied to other organisms.

This dissertation research has provided systems-level insights into the biology of APA in a living organism for the first time, showing that APA is largely regulated at a tissue-specific level and has significant potential to interfere with miRNA induced post-transcriptional gene regulation.

We have developed methods for isolation and sequencing of *C. elegans* tissue-specific mRNA that enabled the study of 3′UTR isoform expression dynamics in an array of somatic tissue types ranging from large muscles to small neurons, to the morphologically dynamic epidermal tissues. This work uncovered extensive APA among *C. elegans* tissues where each tissue commonly prefers selective expression of one 3′UTR isoform, presumably to regulate genes on a tissue-specific basis. This data confirms observations made in mammalian model systems and expands these findings to a living organism with in-tact tissues where developmental contexts are in place.

We identified a significant enrichment of miRNA targets in genes that express tissue-specific 3′UTR isoforms, leading to the hypothesis that these APA events allow many genes to counteract regulation by miRNAs. We were able to genetically validate a tissue-specific APA event in the *C. elegans* body muscle that allows the genes *rack-1* and *tct-1* to escape the broadly expressed miRNA *miR-50*. These results elucidate a role for APA in the tissue-specific modulation of miRNA activities. While our experiments have only demonstrated this mechanism for two genes, the widespread nature of APA in *C. elegans* suggests it may be implicated at a much larger level. We have also used a reverse genetics approach to show that *rack-1* and *tct-1* appear to be important for muscle activity, although we cannot yet be certain these genes have a cell autonomous role in the body muscle and further experiments that specifically target these genes in the body muscle will be required to delineate their roles. It will also be important to establish a biological role for APA using more direct approaches. *C. elegans* is a tractable model for such experiments, since one could use genome editing techniques to modify PAS sites in the genome and score phenotypes that result from the artificial APA events.

Since *rack-1*, *tct-1* and *miR-50* are broadly and co-expressed we hypothesized that many other genes that are ubiquitously expressed among tissues may use APA to allow miRNAs to target distinct sets of genes in each. A study of genes that are commonly expressed among eight total somatic worm tissues uncovered extensive APA among these genes. Commonly expressed genes were also enriched in predicted miRNA targets. We found that almost half of the predicted miRNA targets in this gene pool are lost to APA-induced 3′UTR shortening events. These data agree with studies from human cell lines and suggest that commonly transcribed genes may more often control their dosage at the post-transcriptional level. Former transcriptome studies that have only examined transcript levels, for example by using microarray-based technologies, may be misleading since APA appears to extensively effect how these genes interface with miRNA targeting events among cell types. This dissertation research further improves upon past data as it provides evidence for this mechanism in a living organism that can be genetically manipulated to further study its precise implications.

How the cleavage and polyadenylation machinery discriminate between multiple PAS sites in the same 3′UTR is a long-standing mystery. We detected extensive tissue-specific APA leading to the hypothesis that tissue-specific *cis*-acting elements near the cleavage site may play a role in driving these decisions. We therefore closely examined the sequence regions near PAS sites in muscle and intestine tissues for motifs, including

PAS element usage, that may play a mechanistic role in tissue-specific APA. This analysis uncovered only minor changes in overall PAS element usage and nearly identical nucleotide composition in sequence regions near the cleavage sites, arguing against a model where such tissue-specific sequence elements are major factors controlling APA. Future studies will need to further examine this hypothesis by extending the analysis of the sequence regions to the five additional tissues we have now profiled. Additional work will also be required to now address the alternate hypotheses that tissue-specific factors accessory to the core 3′end formation complex are involved in these mechanisms. The research presented here provides a platform for this future work, since it is now known exactly where many 3′UTR isoforms are expressed and a pool of potential tissue-specific factors is already defined by our tissue-specific expression data. Further, high-throughput genetic screens for such factors are feasible in worms where experiments can be performed *in vivo* where developmental contexts are in place.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Figure 5.2. Alternative polyadenylation as a novel, unexplored mode of post-transcriptional gene regulation. This illustration depicts the major modes of regulation (light gray circles) that precisely dose gene expression at each step of the central dogma (dark gray boxes), eventually culminating in tissue identity. miRNAs are well known for their role in repressing gene expression at the post-transcriptional level. Current data suggests APA is a key player in the same mode of gene regulation that interferes with miRNA regulation.

The RNA research field exploded following the discovery of miRNAs. Over the last several decades, their pervasive roles in normal cellular states, through development, and in disease have been intensively studied and characterized (**Figure 5.2**). It is now becoming clear that miRNA expression patterns do not entirely explain their regulatory activities as APA makes many genes into moving-targets. While still poorly understood, APA is now considered as a mediator of positive regulatory networks that allow genes to escape modes of post-transcriptional gene regulation (**Figure 5.2**). This dissertation research provides clear evidence that 3′UTR heterogeneity driven by APA is largely a tissue-specific phenomenon that appears to have broad impacts on their miRNA targeting networks. We have provided a framework for future studies in this model system that will more precisely delineate how cells use APA to fine-tune their gene expression.

REFERENCES

1. Hales, K.G., et al., *Genetics on the Fly: A Primer on the Drosophila Model System.* Genetics, 2015. **201**(3): p. 815-42.

2. Brenner, S., *The genetics of Caenorhabditis elegans.* Genetics, 1974. **77**(1): p. 71-94.

3. Corsi, A.K., B. Wightman, and M. Chalfie, *A Transparent Window into Biology: A Primer on Caenorhabditis elegans.* Genetics, 2015. **200**(2): p. 387-407.

4. Kumar, S., et al., *Toward 959 nematode genomes.* Worm, 2012. **1**(1): p. 42-50.

5. Saenz-Narciso, B., et al., *The embryonic cell lineage of Caenorhabditis elegans: A modern hieroglyph: The best way to acquire knowledge in Developmental Biology is to learn how this knowledge was derived.* Bioessays, 2015. **37**(3): p. 237-9.

6. Chalfie, M., H.R. Horvitz, and J.E. Sulston, *Mutations that lead to reiterations in the cell lineages of C. elegans.* Cell, 1981. **24**(1): p. 59-69.

7. Sternberg, P.W. and H.R. Horvitz, *The genetic control of cell lineage during nematode development.* Annu Rev Genet, 1984. **18**: p. 489-524.

8. Sternberg, P.W. and H.R. Horvitz, *Pattern formation during vulval development in C. elegans.* Cell, 1986. **44**(5): p. 761-72.

9. Mango, S.E., E.J. Lambie, and J. Kimble, *The pha-4 gene is required to generate the pharyngeal primordium of Caenorhabditis elegans.* Development, 1994. **120**(10): p. 3019-31.

10. Chow, K.L. and S.W. Emmons, *HOM-C/Hox genes and four interacting loci determine the morphogenetic properties of single cells in the nematode male tail.* Development, 1994. **120**(9): p. 2579-92.

11. Chang, B.Y., R.A. Harte, and C.A. Cartwright, *RACK1: a novel substrate for the Src protein-tyrosine kinase.* Oncogene, 2002. **21**(50): p. 7619-29.

12. Hesp, K., G. Smant, and J.E. Kammenga, *Caenorhabditis elegans DAF-16/FOXO transcription factor and its mammalian homologs associate with age-related disease.* Exp Gerontol, 2015. **72**: p. 1-7.

13. Potts, M.B. and S. Cameron, *Cell lineage and cell death: Caenorhabditis elegans and cancer research.* Nat Rev Cancer, 2011. **11**(1): p. 50-8.

14. Herculano-Houzel, S., *The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost.* Proc Natl Acad Sci U S A, 2012. **109 Suppl 1**: p. 10661-8.

15. Towlson, E.K., et al., *The rich club of the C. elegans neuronal connectome.* J Neurosci, 2013. **33**(15): p. 6380-7.

16. White, J.G., et al., *The structure of the nervous system of the nematode Caenorhabditis elegans.* Philos Trans R Soc Lond B Biol Sci, 1986. **314**(1165): p. 1-340.

17. Jin, Y., *Unraveling the mechanisms of synapse formation and axon regeneration: the awesome power of C. elegans genetics.* Sci China Life Sci, 2015. **58**(11): p. 1084-8.

18. Nudel, U., K. Robzyk, and D. Yaffe, *Expression of the putative Duchenne muscular dystrophy gene in differentiated myogenic cell cultures and in the brain.* Nature, 1988. **331**(6157): p. 635-8.

19. Towers, P.R., et al., *Gene expression profiling studies on Caenorhabditis elegans dystrophin mutants dys-1(cx-35) and dys-1(cx18).* Genomics, 2006. **88**(5): p. 642-9.

20. Saito, R.M. and S. van den Heuvel, *Malignant worms: what cancer research can learn from C. elegans.* Cancer Invest, 2002. **20**(2): p. 264-75.

21. Malin, J.Z. and S. Shaham, *Cell Death in C. elegans Development.* Curr Top Dev Biol, 2015. **114**: p. 1-42.

22. Gartner, A., et al., *A conserved checkpoint pathway mediates DNA damage--induced apoptosis and cell cycle arrest in C. elegans.* Mol Cell, 2000. **5**(3): p. 435-43.

23. Sternberg, P.W. and M. Han, *Genetics of RAS signaling in C. elegans.* Trends Genet, 1998. **14**(11): p. 466-72.

24. Zhong, X., et al., *Identification of microRNAs regulating reprogramming factor LIN28 in embryonic stem cells and cancer cells.* J Biol Chem, 2010. **285**(53): p. 41961-71.

25. Galupa, R. and E. Heard, *X-chromosome inactivation: new insights into cis and trans regulation.* Curr Opin Genet Dev, 2015. **31**: p. 57-66.

26. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control.* Nat Rev Genet, 2012. **13**(9): p. 613-26.

27. Reinke, V., M. Krause, and P. Okkema, *Transcriptional regulation of gene expression in C. elegans.* WormBook, 2013: p. 1-34.

28. Lemons, D. and W. McGinnis, *Genomic evolution of Hox gene clusters.* Science, 2006. **313**(5795): p. 1918-22.

29. Herman, M.A., *Hermaphrodite cell-fate specification.* WormBook, 2006: p. 1-16.

30. Cowing, D. and C. Kenyon, *Correct Hox gene expression established independently of position in Caenorhabditis elegans.* Nature, 1996. **382**(6589): p. 353-6.

31. Moss, E.G., *Heterochronic genes and the nature of developmental time.* Curr Biol, 2007. **17**(11): p. R425-34.

32. Kruiswijk, F., C.F. Labuschagne, and K.H. Vousden, *p53 in survival, death and metabolic health: a lifeguard with a licence to kill.* Nat Rev Mol Cell Biol, 2015. **16**(7): p. 393-405.

33. McCarthy, N., *Metastasis: understanding the prowess of mutant p53.* Nat Rev Cancer, 2014. **14**(6): p. 385.

34. Yilmaz, A. and E. Grotewold, *Components and mechanisms of regulation of gene expression.* Methods Mol Biol, 2010. **674**: p. 23-32.

35. Halbeisen, R.E., et al., *Post-transcriptional gene regulation: from genome-wide studies to principles.* Cell Mol Life Sci, 2008. **65**(5): p. 798-813.

36. Jin, L., et al., *Sequestration of mRNAs Modulates the Timing of Translation during Meiosis in Budding Yeast.* Mol Cell Biol, 2015. **35**(20): p. 3448-58.

37. LeGendre, J.B., et al., *RNA targets and specificity of Staufen, a double-stranded RNA-binding protein in Caenorhabditis elegans.* J Biol Chem, 2013. **288**(4): p. 2532-45.

38. Brennan, C.M. and J.A. Steitz, *HuR and mRNA stability.* Cell Mol Life Sci, 2001. **58**(2): p. 266-77.

39. Houseley, J. and D. Tollervey, *The many pathways of RNA degradation.* Cell, 2009. **136**(4): p. 763-76.

40. Glisovic, T., et al., *RNA-binding proteins and post-transcriptional gene regulation.* FEBS Lett, 2008. **582**(14): p. 1977-86.

41. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.

42. Wightman, B., I. Ha, and G. Ruvkun, *Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans.* Cell, 1993. **75**(5): p. 855-62.

43. Reinhart, B.J., et al., *The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.* Nature, 2000. **403**(6772): p. 901-6.

44. Pasquinelli, A.E., et al., *Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA.* Nature, 2000. **408**(6808): p. 86-9.

45. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.

46. Berezikov, E., *Evolution of microRNA diversity and regulation in animals.* Nat Rev Genet, 2011. **12**(12): p. 846-60.

47. Finnegan, E.F. and A.E. Pasquinelli, *MicroRNA biogenesis: regulating the regulators.* Crit Rev Biochem Mol Biol, 2013. **48**(1): p. 51-68.

48. Djuranovic, S., A. Nahvi, and R. Green, *A parsimonious model for gene regulation by miRNAs.* Science, 2011. **331**(6017): p. 550-3.

49. Baek, D., et al., *The impact of microRNAs on protein output.* Nature, 2008. **455**(7209): p. 64-71.

50. Eichhorn, S.W., et al., *mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues.* Mol Cell, 2014. **56**(1): p. 104-15.

51. Fabian, M.R., et al., *Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation.* Mol Cell, 2009. **35**(6): p. 868-80.

52. Hausser, J. and M. Zavolan, *Identification and consequences of miRNA-target interactions--beyond repression of gene expression.* Nat Rev Genet, 2014. **15**(9): p. 599-612.

53. Wolter, J.M., et al., *3'LIFE: a functional assay to detect miRNA targets in high-throughput.* Nucleic Acids Res, 2014. **42**(17): p. e132.

54. Kotagama, K., et al., *A human 3'UTR clone collection to study post-transcriptional gene regulation.* BMC Genomics, 2015. **16**: p. 1036.

55. Stefani, G. and F.J. Slack, *Small non-coding RNAs in animal development.* Nat Rev Mol Cell Biol, 2008. **9**(3): p. 219-30.

56. Reczko, M., et al., *Accurate microRNA Target Prediction Using Detailed Binding Site Accessibility and Machine Learning on Proteomics Data.* Front Genet, 2011. **2**: p. 103.

57. Lall, S., et al., *A genome-wide map of conserved microRNA targets in C. elegans.* Curr Biol, 2006. **16**(5): p. 460-71.

58. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.

59. Ebert, M.S. and P.A. Sharp, *Roles for microRNAs in conferring robustness to biological processes.* Cell, 2012. **149**(3): p. 515-24.

60. Stark, A., et al., *Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution.* Cell, 2005. **123**(6): p. 1133-46.

61. Kanellopoulou, C., et al., *Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing.* Genes Dev, 2005. **19**(4): p. 489-501.

62. Miska, E.A., et al., *Most Caenorhabditis elegans microRNAs are individually not essential for development or viability.* PLoS Genet, 2007. **3**(12): p. e215.

63. Yoo, A.S. and I. Greenwald, *LIN-12/Notch activation leads to microRNA-mediated down-regulation of Vav in C. elegans.* Science, 2005. **310**(5752): p. 1330-3.

64. Cochella, L. and O. Hobert, *Embryonic priming of a miRNA locus predetermines postmitotic neuronal left/right asymmetry in C. elegans.* Cell, 2012. **151**(6): p. 1229-42.

65. Proudfoot, N.J. and J.I. Longley, *The 3' terminal sequences of human alpha and beta globin messenger RNAs: comparison with rabbit globin messenger RNA.* Cell, 1976. **9**(4 PT 2): p. 733-46.

66. Dvir, A., J.W. Conaway, and R.C. Conaway, *Mechanism of transcription initiation and promoter escape by RNA polymerase II.* Curr Opin Genet Dev, 2001. **11**(2): p. 209-14.

67. Sims, R.J., 3rd, R. Belotserkovskaya, and D. Reinberg, *Elongation by RNA polymerase II: the short and long of it.* Genes Dev, 2004. **18**(20): p. 2437-68.

68. Richard, P. and J.L. Manley, *Transcription termination by nuclear RNA polymerases.* Genes Dev, 2009. **23**(11): p. 1247-69.

69. Kuehner, J.N., E.L. Pearson, and C. Moore, *Unravelling the means to an end: RNA polymerase II transcription termination.* Nat Rev Mol Cell Biol, 2011. **12**(5): p. 283-94.

70. Connelly, S. and J.L. Manley, *A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II.* Genes Dev, 1988. **2**(4): p. 440-52.

71. Logan, J., et al., *A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene.* Proc Natl Acad Sci U S A, 1987. **84**(23): p. 8306-10.

72. Luo, W., A.W. Johnson, and D.L. Bentley, *The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model.* Genes Dev, 2006. **20**(8): p. 954-65.

73. Proudfoot, N.J. and G.G. Brownlee, *3' non-coding region sequences in eukaryotic messenger RNA.* Nature, 1976. **263**(5574): p. 211-4.

74. Fitzgerald, M. and T. Shenk, *The sequence 5'-AAUAAA-3'forms parts of the recognition site for polyadenylation of late SV40 mRNAs.* Cell, 1981. **24**(1): p. 251-60.

75. Brockman, J.M., et al., *PACdb: PolyA Cleavage Site and 3'-UTR Database.* Bioinformatics, 2005. **21**(18): p. 3691-3.

76. Gil, A. and N.J. Proudfoot, *Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation.* Cell, 1987. **49**(3): p. 399-406.

77. Murthy, K.G. and J.L. Manley, *The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation.* Genes Dev, 1995. **9**(21): p. 2672-83.

78. Takagaki, Y., L.C. Ryner, and J.L. Manley, *Four factors are required for 3'-end cleavage of pre-mRNAs.* Genes Dev, 1989. **3**(11): p. 1711-24.

79. Takagaki, Y., et al., *A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs.* Genes Dev, 1990. **4**(12A): p. 2112-20.

80. Proudfoot, N.J., *Ending the message: poly(A) signals then and now.* Genes Dev, 2011. **25**(17): p. 1770-82.

81. Takagaki, Y. and J.L. Manley, *Complex protein interactions within the human polyadenylation machinery identify a novel component.* Mol Cell Biol, 2000. **20**(5): p. 1515-25.

82. Lutz, C.S., et al., *Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro.* Genes Dev, 1996. **10**(3): p. 325-37.

83. Shi, Y., et al., *Molecular architecture of the human pre-mRNA 3' processing complex.* Mol Cell, 2009. **33**(3): p. 365-76.

84. Shaye, D.D. and I. Greenwald, *OrthoList: a compendium of C. elegans genes with human orthologs.* PLoS One, 2011. **6**(5): p. e20085.

85. Gupta, I., et al., *Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions.* Mol Syst Biol, 2014. **10**: p. 719.

86. Sherstnev, A., et al., *Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation.* Nat Struct Mol Biol, 2012. **19**(8): p. 845-52.

87. Shepard, P.J., et al., *Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq.* RNA, 2011. **17**(4): p. 761-72.

88. Mangone, M., et al., *The landscape of C. elegans 3'UTRs.* Science, 2010. **329**(5990): p. 432-5.

89. Jan, C.H., et al., *Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs.* Nature, 2011. **469**(7328): p. 97-101.

90. Jenal, M., et al., *The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites.* Cell, 2012. **149**(3): p. 538-53.

91. Simonelig, M., *PABPN1 shuts down alternative poly(A) sites.* Cell Res, 2012. **22**(10): p. 1419-21.

92. Ichinose, J., et al., *Alternative polyadenylation is associated with lower expression of PABPN1 and poor prognosis in non-small cell lung cancer.* Cancer Sci, 2014. **105**(9): p. 1135-41.

93. Berg, M.G., et al., *U1 snRNP determines mRNA length and regulates isoform expression.* Cell, 2012. **150**(1): p. 53-64.

94. Li, W., et al., *Systematic profiling of poly(A)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation.* PLoS Genet, 2015. **11**(4): p. e1005166.

95. Sandberg, R., et al., *Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites.* Science, 2008. **320**(5883): p. 1643-7.

96. Mayr, C. and D.P. Bartel, *Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.* Cell, 2009. **138**(4): p. 673-84.

97. Boutet, S.C., et al., *Alternative polyadenylation mediates microRNA regulation of muscle stem cell function.* Cell Stem Cell, 2012. **10**(3): p. 327-36.

98. Miura, P., et al., *Widespread and extensive lengthening of 3' UTRs in the mammalian brain.* Genome Res, 2013. **23**(5): p. 812-25.

99. Zhang, H., J.Y. Lee, and B. Tian, *Biased alternative polyadenylation in human tissues.* Genome Biol, 2005. **6**(12): p. R100.

100. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

101. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.

102. Sulston, J.E., et al., *The embryonic cell lineage of the nematode Caenorhabditis elegans.* Dev Biol, 1983. **100**(1): p. 64-119.

103. Dupuy, D., et al., *A first version of the Caenorhabditis elegans Promoterome.* Genome Res, 2004. **14**(10B): p. 2169-75.

104. Ramani, A.K., et al., *Genome-wide analysis of alternative splicing in Caenorhabditis elegans.* Genome Res, 2011. **21**(2): p. 342-8.

105. Leung, B., G.J. Hermann, and J.R. Priess, *Organogenesis of the Caenorhabditis elegans intestine.* Dev Biol, 1999. **216**(1): p. 114-34.

106. Mango, S.E., *The C. elegans pharynx: a model for organogenesis.* WormBook, 2007: p. 1-26.

107. Grishkevich, V., T. Hashimshony, and I. Yanai, *Core promoter T-blocks correlate with gene expression levels in C. elegans.* Genome Res, 2011. **21**(5): p. 707-17.

108. Burghoorn, J., et al., *The in vivo dissection of direct RFX-target gene promoters in C. elegans reveals a novel cis-regulatory element, the C-box.* Dev Biol, 2012. **368**(2): p. 415-26.

109. Britton, C., et al., *Identification of promoter elements of parasite nematode genes in transgenic Caenorhabditis elegans.* Mol Biochem Parasitol, 1999. **103**(2): p. 171-81.

110. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.

111. Lianoglou, S., et al., *Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression.* Genes Dev, 2013. **27**(21): p. 2380-96.

112. Colgan, D.F. and J.L. Manley, *Mechanism and regulation of mRNA polyadenylation.* Genes Dev, 1997. **11**(21): p. 2755-66.

113. Wallace, A.M., et al., *Two distinct forms of the 64,000 Mr protein of the cleavage stimulation factor are expressed in mouse male germ cells.* Proc Natl Acad Sci U S A, 1999. **96**(12): p. 6763-8.

114. Shankarling, G.S., et al., *A family of splice variants of CstF-64 expressed in vertebrate nervous systems.* BMC Mol Biol, 2009. **10**: p. 22.

115. Takagaki, Y., et al., *The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation.* Cell, 1996. **87**(5): p. 941-52.

116. de Klerk, E., et al., *Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation.* Nucleic Acids Res, 2012. **40**(18): p. 9089-101.

117. Martin, G., et al., *Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length.* Cell Rep, 2012. **1**(6): p. 753-63.

118. Fox, R.M., et al., *A gene expression fingerprint of C. elegans embryonic motor neurons.* BMC Genomics, 2005. **6**: p. 42.

119. Haenni, S., et al., *Analysis of C. elegans intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3'-end-seq.* Nucleic Acids Res, 2012. **40**(13): p. 6304-18.

120. Steiner, F.A., et al., *Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling.* Genome Res, 2012. **22**(4): p. 766-77.

121. Roy, P.J., et al., *Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans.* Nature, 2002. **418**(6901): p. 975-9.

122. Fox, R.M., et al., *The embryonic muscle transcriptome of Caenorhabditis elegans.* Genome Biol, 2007. **8**(9): p. R188.

123. Gorrepati, L., K.W. Thompson, and D.M. Eisenmann, *C. elegans GATA factors EGL-18 and ELT-6 function downstream of Wnt signaling to maintain the progenitor fate during larval asymmetric divisions of the seam cells.* Development, 2013. **140**(10): p. 2093-102.

124. Pauli, F., et al., *Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans.* Development, 2006. **133**(2): p. 287-95.

125. Van Nostrand, E.L. and S.K. Kim, *Seeing elegance in gene regulatory networks of the worm.* Curr Opin Genet Dev, 2011. **21**(6): p. 776-86.

126. Takayama, J., et al., *Single-cell transcriptional analysis of taste sensory neuron pair in Caenorhabditis elegans.* Nucleic Acids Res, 2010. **38**(1): p. 131-42.

127. Zisoulis, D.G., et al., *Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans.* Nat Struct Mol Biol, 2010. **17**(2): p. 173-9.

128. Blazie, S.M., et al., *Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in Caenorhabditis elegans intestine and muscles.* BMC Biol, 2015. **13**: p. 4.

129. Von Stetina, S.E., et al., *Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C. elegans nervous system.* Genome Biol, 2007. **8**(7): p. R135.

130. Spencer, W.C., et al., *A spatial and temporal map of C. elegans gene expression.* Genome Res, 2011. **21**(2): p. 325-41.

131. Frokjaer-Jensen, C., et al., *Single-copy insertion of transgenes in Caenorhabditis elegans.* Nat Genet, 2008. **40**(11): p. 1375-83.

132. Kurn, N., et al., *Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications.* Clin Chem, 2005. **51**(10): p. 1973-81.

133. Coburn, C. and D. Gems, *The mysterious case of the C. elegans gut granule: death fluorescence, anthranilic acid and the kynurenine pathway.* Front Genet, 2013. **4**: p. 151.

134. McGhee, J.D., *The C. elegans intestine.* WormBook, 2007: p. 1-36.

135. Pukkila-Worley, R. and F.M. Ausubel, *Immune defense mechanisms in the Caenorhabditis elegans intestinal epithelium.* Curr Opin Immunol, 2012. **24**(1): p. 3-9.

136. McGhee, J.D., et al., *The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine.* Dev Biol, 2007. **302**(2): p. 627-45.

137. Reece-Hoyes, J.S., et al., *A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks.* Genome Biol, 2005. **6**(13): p. R110.

138. Haerty, W., et al., *Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution.* BMC Genomics, 2008. **9**: p. 399.

139. Watson, E. and A.J. Walhout, *Caenorhabditis elegans metabolic gene regulatory networks govern the cellular economy.* Trends Endocrinol Metab, 2014.

140. Song, B.M. and L. Avery, *The pharynx of the nematode C. elegans: A model system for the study of motor control.* Worm, 2013. **2**(1): p. e21833.

141. Mango, S.E., *The molecular basis of organ formation: insights from the C. elegans foregut.* Annu Rev Cell Dev Biol, 2009. **25**: p. 597-628.

142. Stein, L., et al., *WormBase: network access to the genome and biology of Caenorhabditis elegans.* Nucleic Acids Res, 2001. **29**(1): p. 82-6.

143. Fukushige, T., et al., *Defining the transcriptional redundancy of early bodywall muscle development in C. elegans: evidence for a unified theory of animal muscle development.* Genes Dev, 2006. **20**(24): p. 3395-406.

144. Chen, L., et al., *Body-wall muscle formation in Caenorhabditis elegans embryos that lack the MyoD homolog hlh-1.* Science, 1992. **256**(5054): p. 240-3.

145. Parker, S., H.S. Peterkin, and H.A. Baylis, *Muscular dystrophy associated mutations in caveolin-1 induce neurotransmission and locomotion defects in Caenorhabditis elegans.* Invert Neurosci, 2007. **7**(3): p. 157-64.

146. Graham, P.L., et al., *Type IV collagen is detectable in most, but not all, basement membranes of Caenorhabditis elegans and assembles on tissues that do not express it.* J Cell Biol, 1997. **137**(5): p. 1171-83.

147. Beermann, M.L., et al., *Prdm1 (Blimp-1) and the expression of fast and slow myosin heavy chain isoforms during avian myogenesis in vitro.* PLoS One, 2010. **5**(4): p. e9951.

148. Yang, C., et al., *Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters.* Gene, 2007. **389**(1): p. 52-65.

149. Neilson, J.R. and R. Sandberg, *Heterogeneity in mammalian RNA 3' end formation.* Exp Cell Res, 2010. **316**(8): p. 1357-64.

150. Hedgecock, E.M. and J.G. White, *Polyploid tissues in the nematode Caenorhabditis elegans.* Dev Biol, 1985. **107**(1): p. 128-33.

151. Gremel, G., et al., *The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling.* J Gastroenterol, 2014.

152. Stoeckius, M., et al., *Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression.* Nat Methods, 2009. **6**(10): p. 745-51.

153. Consortium, C.e.S., *Genome sequence of the nematode C. elegans: a platform for investigating biology.* Science, 1998. **282**(5396): p. 2012-8.

154. Kenigsberg, E. and A. Tanay, *Drosophila functional elements are embedded in structurally constrained sequences.* PLoS Genet, 2013. **9**(5): p. e1003512.

155. van Heeringen, S.J., et al., *Nucleotide composition-linked divergence of vertebrate core promoter architecture.* Genome Res, 2011. **21**(3): p. 410-21.

156. Fenouil, R., et al., *CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters.* Genome Res, 2012. **22**(12): p. 2399-408.

157. Costantini, M., R. Cammarano, and G. Bernardi, *The evolution of isochore patterns in vertebrate genomes.* BMC Genomics, 2009. **10**: p. 146.

158. Fire, A., R. Alcazar, and F. Tan, *Unusual DNA structures associated with germline genetic activity in Caenorhabditis elegans.* Genetics, 2006. **173**(3): p. 1259-73.

159. Kormish, J.D., J. Gaudet, and J.D. McGhee, *Development of the C. elegans digestive tract.* Curr Opin Genet Dev, 2010. **20**(4): p. 346-54.

160. Sugi, T. and Y. Ohtani, *Simplified method for cell-specific gene expression analysis in Caenorhabditis elegans.* Biochem Biophys Res Commun, 2014.

161. Subtelny, A.O., et al., *Poly(A)-tail profiling reveals an embryonic switch in translational control.* Nature, 2014. **508**(7494): p. 66-71.

162. Masamha, C.P., et al., *CFIm25 links alternative polyadenylation to glioblastoma tumour suppression.* Nature, 2014. **510**(7505): p. 412-6.

163. Haider, S. and R. Pal, *Integrated analysis of transcriptomic and proteomic data.* Curr Genomics, 2013. **14**(2): p. 91-110.

164. Ghazalpour, A., et al., *Comparative analysis of proteome and transcriptome variation in mouse.* PLoS Genet, 2011. **7**(6): p. e1001393.

165. Kudlow, B.A., L. Zhang, and M. Han, *Systematic analysis of tissue-restricted miRISCs reveals a broad role for microRNAs in suppressing basal activity of the C. elegans pathogen response.* Mol Cell, 2012. **46**(4): p. 530-41.

166. Seiler, C.Y., et al., *DNASU plasmid and PSI:Biology-Materials repositories: resources to accelerate biological research.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1253-60.

167. Frøkjær-Jensen, http://www.wormbuilder.org.

168. Mello, C.C., et al., *Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences.* EMBO J, 1991. **10**(12): p. 3959-70.

169. Zisoulis, D.G., et al., *Autoregulation of microRNA biogenesis by let-7 and Argonaute.* Nature, 2012. **486**(7404): p. 541-4.

170. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.

171. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

172. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.

173. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.

174. Zeiser, E., et al., *MosSCI and gateway compatible plasmid toolkit for constitutive and inducible expression of transgenes in the C. elegans germline.* PLoS One, 2011. **6**(5): p. e20082.

175. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.

176. Mathelier, A., et al., *JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2014. **42**(Database issue): p. D142-7.

177. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

178. Rajewsky, N., *microRNA target predictions in animals.* Nat Genet, 2006. **38 Suppl**: p. S8-13.

179. Gurtan, A.M. and P.A. Sharp, *The role of miRNAs in regulating gene expression networks.* J Mol Biol, 2013. **425**(19): p. 3582-600.

180. Lim, L.P., et al., *Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.* Nature, 2005. **433**(7027): p. 769-73.

181. Adams, D.R., D. Ron, and P.A. Kiely, *RACK1, A multifaceted scaffolding protein: Structure and function.* Cell Commun Signal, 2011. **9**: p. 22.

182. Nilsson, J., et al., *Regulation of eukaryotic translation by the RACK1 protein: a platform for signalling molecules on the ribosome.* EMBO Rep, 2004. **5**(12): p. 1137-41.

183. Battaini, F. and A. Pascale, *Protein kinase C signal transduction regulation in physiological and pathological aging.* Ann N Y Acad Sci, 2005. **1057**: p. 177-92.

184. Cheng, D., et al., *Receptor for activated protein kinase C1 regulates cell proliferation by modulating calcium signaling.* Hypertension, 2011. **58**(4): p. 689-95.

185. Ohn, T., et al., *A functional RNAi screen links O-GlcNAc modification of ribosomal proteins to stress granule and processing body assembly.* Nat Cell Biol, 2008. **10**(10): p. 1224-31.

186. Chu, Y.D., et al., *RACK-1 regulates let-7 microRNA expression and terminal cell differentiation in Caenorhabditis elegans.* Cell Cycle, 2014. **13**(12): p. 1995-2009.

187. Amson, R.B., et al., *Isolation of 10 differentially expressed cDNAs in p53-induced apoptosis: activation of the vertebrate homologue of the drosophila seven in absentia gene.* Proc Natl Acad Sci U S A, 1996. **93**(9): p. 3953-7.

188. Tuynder, M., et al., *Biological models and genes of tumor reversion: cellular reprogramming through tpt1/TCTP and SIAH-1.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 14976-81.

189. Bommer, U.A., et al., *Growth-factor dependent expression of the translationally controlled tumour protein TCTP is regulated through the PI3-K/Akt/mTORC1 signalling pathway.* Cell Signal, 2015. **27**(8): p. 1557-68.

190. Zarogoulidis, P., et al., *mTOR pathway: A current, up-to-date mini-review (Review).* Oncol Lett, 2014. **8**(6): p. 2367-2370.

191. Maeng, J., et al., *Up-regulation of Rhoa/Rho kinase pathway by translationally controlled tumor protein in vascular smooth muscle cells.* Int J Mol Sci, 2014. **15**(6): p. 10365-76.

192. Meyvis, Y., et al., *Analysis of the translationally controlled tumour protein in the nematodes Ostertagia ostertagi and Caenorhabditis elegans suggests a pivotal role in egg production.* Int J Parasitol, 2009. **39**(11): p. 1205-13.

193. Demarco, R.S. and E.A. Lundquist, *RACK-1 acts with Rac GTPase signaling and UNC-115/abLIM in Caenorhabditis elegans axon pathfinding and cell migration.* PLoS Genet, 2010. **6**(11): p. e1001215.

194. Zorio, D.A., et al., *Operons as a common form of chromosomal organization in C. elegans.* Nature, 1994. **372**(6503): p. 270-2.

195. Mikl, M. and C.R. Cowan, *Alternative 3' UTR selection controls PAR-5 homeostasis and cell polarity in C. elegans embryos.* Cell Rep, 2014. **8**(5): p. 1380-90.

196. Frand, A.R., S. Russel, and G. Ruvkun, *Functional genomic analysis of C. elegans molting.* PLoS Biol, 2005. **3**(10): p. e312.

197. Hartig, S.M., *Basic image analysis and manipulation in ImageJ.* Curr Protoc Mol Biol, 2013. **Chapter 14**: p. Unit14 15.

198. Kamath, R.S., et al., *Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.* Nature, 2003. **421**(6920): p. 231-7.

199. Timmons, L., D.L. Court, and A. Fire, *Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans.* Gene, 2001. **263**(1-2): p. 103-12.

200. Matoulkova, E., et al., *The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells.* RNA Biol, 2012. **9**(5): p. 563-76.

201. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.

202. Stinchcomb, D.T., et al., *Extrachromosomal DNA transformation of Caenorhabditis elegans.* Mol Cell Biol, 1985. **5**(12): p. 3484-96.

203. McCue, H.V., et al., *Expression profile of a Caenorhabditis elegans model of adult neuronal ceroid lipofuscinosis reveals down regulation of ubiquitin E3 ligase components.* Sci Rep, 2015. **5**: p. 14392.

204. Horn, M., et al., *DRE-1/FBXO11-dependent degradation of BLMP-1/BLIMP-1 governs C. elegans developmental timing and maturation.* Dev Cell, 2014. **28**(6): p. 697-710.

205. Diaz, V.M. and A.G. de Herreros, *F-box proteins: Keeping the epithelial-to-mesenchymal transition (EMT) in check.* Semin Cancer Biol, 2016. **36**: p. 71-9.

206. Ayyadevara, S., et al., *Caenorhabditis elegans PI3K mutants reveal novel genes underlying exceptional stress resistance and lifespan.* Aging Cell, 2009. **8**(6): p. 706-25.

207. Seifert, M., E. Schmidt, and R. Baumeister, *The genetics of synapse formation and function in Caenorhabditis elegans.* Cell Tissue Res, 2006. **326**(2): p. 273-85.

208. McGrath, P.T., et al., *Parallel evolution of domesticated Caenorhabditis species targets pheromone receptor genes.* Nature, 2011. **477**(7364): p. 321-5.

209. Schuske, K., A.A. Beg, and E.M. Jorgensen, *The GABA nervous system in C. elegans.* Trends Neurosci, 2004. **27**(7): p. 407-14.

210. McIntire, S.L., et al., *The GABAergic nervous system of Caenorhabditis elegans.* Nature, 1993. **364**(6435): p. 337-41.

211. Cinar, H., S. Keles, and Y. Jin, *Expression profiling of GABAergic motor neurons in Caenorhabditis elegans.* Curr Biol, 2005. **15**(4): p. 340-6.

212. Ferguson, E.L. and H.R. Horvitz, *The multivulva phenotype of certain Caenorhabditis elegans mutants results from defects in two functionally redundant pathways.* Genetics, 1989. **123**(1): p. 109-21.

213. Brockie, P.J., et al., *The C. elegans glutamate receptor subunit NMR-1 is required for slow NMDA-activated currents that regulate reversal frequency during locomotion.* Neuron, 2001. **31**(4): p. 617-30.

214. de Kok, Y.J., et al., *Association between X-linked mixed deafness and mutations in the POU domain gene POU3F4.* Science, 1995. **267**(5198): p. 685-8.

215. Janic, A., et al., *Ectopic expression of germline genes drives malignant brain tumor growth in Drosophila.* Science, 2010. **330**(6012): p. 1824-7.

216. Pierce, S.B., et al., *Regulation of DAF-2 receptor signaling by human insulin and ins-1, a member of the unusually large and diverse C. elegans insulin gene family.* Genes Dev, 2001. **15**(6): p. 672-86.

217. Pocock, R., et al., *Neuronal function of Tbx20 conserved from nematodes to vertebrates.* Dev Biol, 2008. **317**(2): p. 671-85.

218. Jafari, G., et al., *The UNC-4 homeobox protein represses mab-9 expression in DA motor neurons in Caenorhabditis elegans.* Mech Dev, 2011. **128**(1-2): p. 49-58.

219. Portereiko, M.F. and S.E. Mango, *Early morphogenesis of the Caenorhabditis elegans pharynx.* Dev Biol, 2001. **233**(2): p. 482-94.

220. Rasmussen, J.P., et al., *Cell interactions and patterned intercalations shape and link epithelial tubes in C. elegans.* PLoS Genet, 2013. **9**(9): p. e1003772.

221. Chisholm, A.D. and J. Hardin, *Epidermal morphogenesis.* WormBook, 2005: p. 1-22.

222. Jackson, B.M., et al., *Use of an activated beta-catenin to identify Wnt pathway target genes in caenorhabditis elegans, including a subset of collagen genes expressed in late larval development.* G3 (Bethesda), 2014. **4**(4): p. 733-47.

223. Hao, L., et al., *Comprehensive analysis of gene expression patterns of hedgehog-related genes.* BMC Genomics, 2006. **7**: p. 280.

224. Cassata, G., et al., *ceh-16/engrailed patterns the embryonic epidermis of Caenorhabditis elegans.* Development, 2005. **132**(4): p. 739-49.

225. Kouns, N.A., et al., *NHR-23 dependent collagen and hedgehog-related genes required for molting.* Biochem Biophys Res Commun, 2011. **413**(4): p. 515-20.

226. Harandi, O.F. and V.R. Ambros, *Control of stem cell self-renewal and differentiation by the heterochronic genes and the cellular asymmetry machinery in Caenorhabditis elegans.* Proc Natl Acad Sci U S A, 2015. **112**(3): p. E287-96.

227. Joshi, P.M., et al., *Caenorhabditis elegans as a model for stem cell biology.* Dev Dyn, 2010. **239**(5): p. 1539-54.

228. Aspock, G. and T.R. Burglin, *The Caenorhabditis elegans distal-less ortholog ceh-43 is required for development of the anterior hypodermis.* Dev Dyn, 2001. **222**(3): p. 403-9.

229. Hughes, S., et al., *CEH-20/Pbx and UNC-62/Meis function upstream of rnt-1/Runx to regulate asymmetric divisions of the C. elegans stem-like seam cells.* Biol Open, 2013. **2**(7): p. 718-27.

230. Betel, D., et al., *The microRNA.org resource: targets and expression.* Nucleic Acids Res, 2008. **36**(Database issue): p. D149-53.

231. Alvarez-Saavedra, E. and H.R. Horvitz, *Many families of C. elegans microRNAs are not essential for development or viability.* Curr Biol, 2010. **20**(4): p. 367-73.

232. Baugh, L.R. and P.W. Sternberg, *DAF-16/FOXO regulates transcription of cki-1/Cip/Kip and repression of lin-4 during C. elegans L1 arrest.* Curr Biol, 2006. **16**(8): p. 780-5.

233. Martinez, N.J., et al., *Genome-scale spatiotemporal analysis of Caenorhabditis elegans microRNA promoter activity.* Genome Res, 2008. **18**(12): p. 2005-15.

234. Simon, D.J., et al., *The microRNA miR-1 regulates a MEF-2-dependent retrograde signal at neuromuscular junctions.* Cell, 2008. **133**(5): p. 903-15.

235. Hoffman, Y., et al., *3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells.* PLoS Genet, 2016. **12**(2): p. e1005879.

236. Kunitomo, H., et al., *Identification of ciliated sensory neuron-expressed genes in Caenorhabditis elegans using targeted pull-down of poly(A) tails.* Genome Biol, 2005. **6**(2): p. R17.

237. Blazie, S.M., et al., *Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in Caenorhabditis elegans intestine and muscles.* BMC Biol, 2015. **13**(1): p. 4.

238. Diederichs, S., et al., *The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations.* EMBO Mol Med, 2016.

239. Chen, C.Y., et al., *Lengthening of 3'UTR increases with morphological complexity in animal evolution.* Bioinformatics, 2012. **28**(24): p. 3178-81.

240. Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.

241. Nam, J.W., et al., *Global analyses of the effect of different cellular contexts on microRNA targeting.* Mol Cell, 2014. **53**(6): p. 1031-43.

242. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

243. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

244. Hubbard, T., et al., *The Ensembl genome database project.* Nucleic Acids Res, 2002. **30**(1): p. 38-41.